DOCUMENT RESUME

ED 399 524 CS 012 607

AUTHOR	Abdullah, Mardziah Hayati, Comp.
TITLE	Standardized and Alternative Assessment. Hot Topic Guide 59.
INSTITUTION	Indiana Univ., Bloomington. School of Education.
PUB DATE	96
NOTE	207p.
PUB TYPE	Information Analyses (070) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Annotated Bibliographies; Class Activities; Elementary Secondary Education; Liberal Arts; Performance Based Assessment; *Standardized Tests; *Student Evaluation; *Testing; Workshops
IDENTIFIERS	*Alternative Assessment; Authentic Assessment

ABSTRACT

One of a series of educational packages designed for implementation either in a workshop atmosphere or through individual study, this Hot Topic guide presents a variety of materials designed to assist educators in designing and implementing classroom projects and activities centering on the topic of standardized and alternative assessment. The Hot Topic guide contains guidelines for workshop use; an overview of standardized and alternative assessment; and seven articles (from scholarly and professional journals) and ERIC documents on the topic. A 19-item annotated bibliography of items in the ERIC database on the topic is attached. (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 399 524

HOT TOPIC
GUIDE 59
Standardized and Alternative Assessment

TABLE OF CONTENTS:

HELPFUL GUIDELINES FOR WORKSHOP USE
Suggestions for using this Hot Topic Guide as a professional development tool.

OVERVIEW/LECTURE
Standardized and Alternative Assessment
by Mardziah Hayati Abdullah

ARTICLES AND ERIC DOCUMENTS
- The Need for a New Science of Assessment
- Raising Standardized Test Scores and the Origins of Test Score Pollution
- Performance-Based Assessment and Educational Equity
- Assessment and the Morality of Testing
- Assessment Worthy of the Liberal Arts
- The Morality of Test Security
- Testing and Tact

BIBLIOGRAPHY
A collection of selected references and abstracts obtained directly from the ERIC database.

Indiana University, Bloomington. School of Education.
Compiler: Mardziah Hayati Abdullah
Series Editors: Carl Smith, Eleanor Macfarlane, and Christopher Essex

1996

CS01260 7

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
C Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

In-Service Workshops and Seminars:
Suggestions for Using this Hot Topic Guide as a Professional Development Tool

Before the Workshop:

- Carefully review the materials presented in this Hot Topic Guide. Think about how these concepts and projects might be applied to your particular school or district.
- As particular concepts begin to stand out in your mind as being important, use the Bibliography section (found at the end of the packet) to seek out additional resources dealing specifically with those concepts.
- Look over the names of the teachers and researchers who wrote the packet articles and/or are listed in the Bibliography. Are any of the names familiar to you? Do any of them work in your geographical area? Do you have colleagues or acquaintances who are engaged in similar research and/or teaching? Perhaps you could enlist their help and expertise as you plan your workshop or seminar.
- As you begin to plan your activities, develop a mental "movie" of what you'd like to see happening in the classroom as a result of this in-service workshop or seminar. Keep this vision in mind as a guide to your planning.

During the Workshop:

- Provide your participants with a solid grasp of the important concepts that you have acquired from your reading, but don't load them down with excessive detail, such as lots of hard-to-remember names, dates or statistics. You may wish to use the Overview/Lecture section of this packet as a guide for your introductory remarks about the topic.
- Try modeling the concepts and teaching strategies related to the topic by "teaching" a minilesson for your group.
- Remember, if your teachers and colleagues ask you challenging or difficult questions about the topic, that they are not trying to discredit you or your ideas. Rather, they are trying to prepare themselves for situations that might arise as they implement these ideas in their own classrooms.
- If any of the participants are already using some of these ideas in their own teaching, encourage them to share their experiences.
- Even though your workshop participants are adults, many of the classroom management principles that you use every day with your students still apply. Workshop participants, admittedly, have a longer attention span and can sit still longer than your second-graders; but not that much longer. Don't have a workshop that is just a "sit down, shut up, and listen" session. Vary the kinds of presentations and activities you provide in your workshops. For instance, try to include at least one hands-on activity so that the participants will begin to get a feel for how they might apply the concepts that you are discussing in your workshop.
- Try to include time in the workshop for the participants to work in small groups. This time may be a good opportunity for them to formulate plans for how they might use the concepts just discussed in their own classrooms.
- Encourage teachers to go "a step further" with what they have learned in the workshop. Provide additional resources for them to continue their research into the topics discussed, such as books, journal articles, Hot Topic Guides, teaching materials, and local experts. Alert them to future workshops/conferences on related topics.

11/94

After the Workshop:

- Follow up on the work you have done. Have your workshop attendees fill out an End-of-Session Evaluation (a sample is included on the next page). Emphasize that their responses are anonymous. The participants' answers to these questions can be very helpful in planning your next workshop. After a reasonable amount of time (say a few months or a semester), contact your workshop attendees and inquire about how they have used, or haven't used, the workshop concepts in their teaching. Have any surprising results come up? Are there any unforeseen problems?
- When teachers are trying the new techniques, suggest that they invite you to observe their classes. As you discover success stories among teachers from your workshop, share them with the other attendees, particularly those who seem reluctant to give the ideas a try.
- Find out what other topics your participants would like to see covered in future workshops and seminars. There are nearly sixty Hot Topic Guides, and more are always being developed. Whatever your focus, there is probably a Hot Topic Guide that can help. An order form follows the table of contents in this packet.

Are You Looking for University Course Credit?

Indiana University's Distance Education program is offering new one-credit-hour Language Arts Education minicourses on these topics:

Elementary:
Language Learning and Development
Varied Writing Strategies
Parents and the Reading Process
Exploring Creative Writing with Elementary Students

Secondary:
Varied Writing Strategies
Thematic Units and Literature
Exploring Creative Writing with Secondary Students

K-12:
Reading across the Curriculum
Writing across the Curriculum
Organization of the Classroom

Course Requirements:
These minicourses are taught by correspondence. Minicourse reading materials consist of Hot Topic Guides and ERIC/EDINFO Press books. You will be asked to write Goal Statements and Reaction Papers for each of the assigned reading materials, and a final Synthesis paper.

I really enjoyed working at my own pace..... It was wonderful to have everything so organized...and taken care of in a manner where I really felt like I was a student, however "distant" I was....."
--Distance Education student

Three-Credit-Hour Courses are also offered (now with optional videos!):
Advanced Study in the Teaching of:

- Reading in the Elementary School
- Language Arts in the Elementary School
- Secondary School English/Language Arts
- Reading in the Secondary School

Writing as a Response to Reading
Developing Parent Involvement Programs
Critical Thinking across the Curriculum
Organization and Administration of a School Reading Program

For More Information:
For course outlines and registration instructions, please contact:

Distance Education Office
Smith Research Center, Suite 150
2805 East 10th Street
Bloomington, IN 47408-2698
1-800-759-4723 or (812) 855-5847

Planning a Workshop Presentation
Worksheet

Major concepts you want to stress in this presentation:

1) _____

2) _____

3) _____

Are there additional resources mentioned in the Bibliography that would be worth locating? Which ones? How could you get them most easily?

_____

_____

Are there resource people available in your area whom you might consult about this topic and/or invite to participate? Who are they?

_____

_____

What would you like to see happen in participants' classrooms as a result of this workshop? Be as specific as possible.

_____

_____

_____

_____

Plans for followup to this workshop: [peer observations, sharing experiences, etc.]

_____

_____

Agenda for Workshop
Planning Sheet

Introduction/Overview:

[What would be the most effective way to present the major concepts that you wish to convey?]

_____

_____

_____

Activities that involve participants and incorporate the main concepts of this workshop:

1) _____

_____

_____

_____

2) _____

_____

_____

_____

Applications:

Encourage participants to plan a mini-lesson for their educational setting that draws on these concepts. [One possibility is to work in small groups, during the workshop, to make a plan and then share it with other participants.]

Your plan to make this happen:

_____

_____

Evaluation:

[Use the form on the next page, or one you design, to get feedback from participants about your presentation.]

End-of-Session Evaluation

Now that today's meeting is over, we would like to know how you feel and what you think about the things we did so that we can make them better. Your opinion is important to us. Please answer all questions honestly. Your answers are confidential.

1. Check (✓) to show if today's meeting was

   ❑ Not worthwhile ❑ Somewhat worthwhile ❑ Very worthwhile

2. Check (✓) to show if today's meeting was

   ❑ Not interesting ❑ Somewhat interesting ❑ Very interesting

3. Check (✓) to show if today's leader was

   ❑ Not very good ❑ Just O.K. ❑ Very good

4. Check (✓) to show if the meeting helped you get any useful ideas about how you can make positive changes in the classroom.

   ❑ Very little ❑ Some ❑ Very much

5. Check (✓) to show if today's meeting was

   ❑ Too long ❑ Too short ❑ Just about right

6. Check (✓) whether you would recommend today's meeting to a colleague.

   ❑ Yes ❑ No

7. Check (✓) to show how useful you found each of the things we did or discussed today.

   Getting information/new ideas.

   ❑ Not useful ❑ Somewhat useful ❑ Very useful

   Seeing and hearing demonstrations of teaching techniques.

   ❑ Not useful ❑ Somewhat useful ❑ Very useful

   Getting materials to read.

   ❑ Not useful ❑ Somewhat useful ❑ Very useful

Listening to other teachers tell about their own experiences.

❑ Not useful ❑ Somewhat useful ❑ Very useful

Working with colleagues in a small group to develop strategies of our own.

❑ Not useful ❑ Somewhat useful ❑ Very useful

Getting support from others in the group.

❑ Not useful ❑ Somewhat useful ❑ Very useful

8. Please write one thing that you thought was best about today:

9. Please write one thing that could have been improved today:

10. What additional information would you have liked?

11. Do you have any questions you would like to ask?

12. What additional comments would you like to make?

Thank you for completing this form.

STANDARDIZED AND ALTERNATIVE ASSESSMENT

Overview by
Mardziah Hayati Abdullah
M.S. Language Education, Indiana University

We evaluate every day. We make judgments about the most common things without even realizing it when we comment on how a dish tastes, which TV program is more entertaining or how to perform a chore better. Understandably, we would expect evaluation to be an important part of a system in which so much national interest and expenditure is invested: education. Evaluation should inform educators about how well they have done or are doing, and it should indicate how to educate better. However, educational assessment has other social and political consequences which significantly impact society and thus provoke debate. This overview introduces some of the major issues surrounding educational assessment and provides a guide to further reading on particular aspects.

Trends in Assessment

The practice of evaluating human ability and performance in an organized manner has been around for centuries, dating as far back as 2000 B.C. when Chinese civil examinations were established in a move toward appointing civil servants based on meritocracy instead of unfair preference. Since then, trends in assessment have undergone changes in response to changes in the philosophy of education, assumptions about learning, and political ideologies.

In the 400's B.C, the Greek philosopher Socrates used oral conversational methods for examining rhetorical abilities in presenting and defending arguments. Much later, in the 1700's, early testing in the United States also involved oral examinations conducted by faculty who used their expert judgment to determine the quality of their students' performances, much like the oral defenses required of doctoral students today. Horace Mann started using written standardized tests in Massachusetts and Connecticut in the 1830's, leading to the development of Mann's Boston Survey in 1846, the first printed test for large-scale assessment of student achievement in various disciplines. The tests were discontinued because the results were not used.

Standardized examinations were again recommended after 1895 when Joseph Rice conducted tests in a number of large school systems that yielded large differences in math scores among schools. In the 1900's Thorndike, often called the father of the educational testing movement, persuaded educators to measure human change. There was a call for immediate, demonstrable results, and by 1915, testing

had become the primary means of evaluating schools. A lack of trust in teachers developed because of the assumption that they were biased in evaluating students, and standardized norm-referenced tests emerged so that individual students' or group scores could be compared against those of other individuals or groups on a bell-shaped normal curve. Dewey's progressive education movement favored education based on natural learning and practical problems in the 1930's, and while that may have called for more authentic testing, standardized norm-referenced testing was still employed, leading to a boom in the technical development of tests to compare schools on a larger scale.

As educators realized that standardized test outcomes did not adequately reflect the complexity of learning, opposition against standardized testing grew. The 1980's saw a rise in the use of criterion-referenced tests that, in reality, are also standardized, but students are assessed against a set of criteria instead of other students. By the 1990's, however, educators were raising and documenting more and more problems with standardized testing. In answer, educators explored alternatives to standardized testing which fall under the umbrella term *alternative assessment*. Different forms of alternative assessment are still being developed and tried enthusiastically by some educators, but they, too, face opposition. In Chapter 1 and part of Chapter 8 in the book Toward a New Science of Educational Testing and Assessment, which have been included in this Hot Topic guide, Berlak discusses some of the epistemological contentions in the history of evaluation.

Testing and assessment

Before engaging in a discussion on standardized testing and alternative assessment, I think it is important to read Chapter 1 in the reading text for this guide, Assessing Student Performance: Exploring the Purpose and Limits of Testing, in which Wiggins (1993) makes an enlightening distinction between *testing* and *assessment*.

In a nutshell, a test is a "one- shot" procedure which assumes to measure the test-taker's ability or knowledge. It requires the test-taker to provide uniform, `correct' responses to items formulated and scored by other parties using criteria over which the test-taker has no say. Thus, Wiggins claims, to "test" a student is "a practice of determining whether the student has mastered what is orthodox" (p. 10). Tests are often "secure" (the contents kept secret) before they are administered. The role of human judgment in scoring is deliberately minimized. Standardized items and scoring also necessarily minimizes responsiveness to individual test-takers and contexts. Assessment, in contrast, is a comprehensive, multi-faceted analysis of performance; it must be judgment-based and personal" (p.13). By Wiggins' definition, assessment requires the systematic collection of data about a student's performance

using multiple and varied techniques, including assessors' observations. Unlike scores obtained from one-shot tests, assessment involves the "integration of (diverse) information in a summary judgment" about a student (p.13) and necessitates the assessor's personal and subjective evaluation of a particular individual's performance in a particular context. The student has more say in his own assessment.

If one accepts the distinctions described above, then testing cannot be assessment, and vice-versa. Wiggins' book clarifies aspects of evaluation that further distinguish the two, such as the practice of assessment (Chapter 2), test security versus tact (Chapter 4) and the nature of feedback (Chapter 6). (Note: Chapter 6 is not a required reading for this Hot Topic guide but is included in the one on Measurement Issues).

Standardized testing

A standardized test is typically one in which the same questions, in the same form, are given to everyone who is being evaluated. The test may be scored by a machine or by different raters following a common rubric. On one hand, standardized testing procedures have made evaluation convenient by offering neatly packaged and often easily obtainable test items. Machine scoring of multiple- choice questions has also made scoring easier and free of human error, and even Wiggins acknowledges that there are virtues in minimizing bias and error in human judgment. However, more and more educators are presenting strong arguments against standardized testing. For instance, bias can also be built into the questions themselves, as test items inevitably reflect particular perspectives that may not be shared by test-takers. Thus, incorrect responses need not indicate erroneous understanding.

Berlak discusses four assumptions underlying what he considers to be the current *psychometric* paradigm of standardized testing, along with counter-arguments that support the emergence of a *contextual* or curriculum-sensitive paradigm. One assumption is that there can be a single, established consensus about what a test score means about individuals all over the world, while the argument is that there are "plural and contradictory" perspectives on what it means to be able or competent in a certain area. Secondly, the psychometric paradigm assumes that `scientific' techniques and instruments are value-neutral and thus should yield objective measurements; this is countered by the assertion that there are different learning situations and culturally ingrained perspectives, making it morally wrong to measure students' performance without considering those factors. Another assumption is that cognitive learning can be separated from affective learning and measured likewise, but the contextual paradigm argues that cognitive ability cannot be assessed independently from other human factors. Finally, the psychometric paradigm sees a

need for centralized control, whereas assessment reform calls for decentralized decision making.

Standardized testing has also raised important issues of equity, deprofessionalization of teachers and the negative use of test scores for purposes such as placement, tracking and grade retention. Pages 5-18 of Linda Darling-Hammond's article Symposium: Equity in Educational Assessment (included in this guide) presents these issues convincingly, highlighting the need to consider the consequential validity of testing. Haladyna, Nolen and Haas, in their 1991 article, Raising Standardized Achievement Scores and the Origins of Test Score Pollution (found in this guide), point out yet another consequence: unethical practices that schools engage in as a result of society's concern with standardized scores as indicators of performance. The escalating academic demand has even led to readiness screening and grade retention being imposed on children as young as kindergartners (Shepard & Smith, 1988). With these and other criticisms being levied against standardized testing, assessment, as Wiggins defines it, is increasingly being sought as an alternative evaluation approach.

Alternative assessment

Alternative assessment has come to be used as an umbrella term that covers authentic assessment and performance assessment, and incorporates techniques such as portfolio assessment, exhibitions of mastery, projects, profiles and discourse assessment.

Authentic assessment refers to the assessment of achievement or ability shown through performances which are modeled on real-life tasks or practices in beyond-school settings. It is performance-based as it evaluates performances such as the ability to integrate, apply or produce knowledge in contexts representative of or similar to real-life situations that (ideally) have personal value for the student. The forms and criteria of assessment are similar to those that might be used in real situations. Thus, authentic assessment is likely to have a combination of some or all of the following characteristics: *ongoing, continuous, personalized, negotiable, with multiple indicators, and possibly conducted by more than one assessor*, although this list is not exhaustive. Fred Newmann and Douglas Archbald discuss criteria for authentic assessment in Chapter 4 of the Berlak (et.al.) book (not included in the guide). In Chapter 7 (also not included), the same writers describe practices for assessing academic achievement that meet one or more of those criteria.

Wiggins' book offers one of the best discussions available on what should constitute *performance assessment*. Gitomer (in press) defines a performance task as "one that simultaneously requires the use of knowledge, skills and values that are

recognized as important in a domain of study and is qualitatively consistent with tasks that members of discipline-based communities might conceivably engage in." The alternative assessment techniques mentioned earlier represent performance tasks that vary in scope of content, method of initiation and method of presentation. Texts by the following (not required reading but listed under references at the end of this overview) provide more detailed descriptions of these techniques: Newmann in Berlak, et.al.(discourse assessment), Mabry (demonstrations of mastery), Stenmark (performance tasks), Archbald and Newmann (portfolios and profiles). (The last text is a required reading for the Hot Topic guide on Measurement Issues.)

How do alternative assessment techniques differ from standardized tests in language education? Discourse assessment, as described by Newmann, is an example: it is the assessment of a student's ability to present coherent, 'whole' (as opposed to fragmented) responses in discursive forms such as narratives or arguments, which reflect synthesized, personal, critical and contextually appropriate use of content and language. It also helps to understand what discourse assessment is *not*, namely, the checking of answers to responses, both of which are presented in a sequence that has no obvious meaning or purpose beyond that of testing discrete items of knowledge. Discourse assessment is *in principle* authentic assessment, as it requires a form of discursive performance that we engage in real-life situations, such as when we write letters, argue to defend our position or choices, negotiate with others or present information. However, authenticity comes only with contextual appropriateness, hence we must consider the nature of the performance task and criteria. For instance, if students in a rural village (who are likely to become the future leaders) were facing the possibility of their only soccer field being used for the construction of teachers' quarters, an authentic assessment task would be for them to present strong oral or written arguments for stopping the housing project. The *assessor* must also be familiar with the discursive elements that are culturally appropriate for that situation, otherwise an inauthentic assessment will be rendered. A standardized test necessarily excludes such personalized attention.

Indeed, Wiggins and Mabry view alternative assessment within the framework of a *personalized* assessment paradigm. Going beyond aligning curriculum with assessment, which is the crux of a contextual paradigm, a personalized paradigm includes student- sensitive content, variable times and settings, greater student selection and the essential feature of self-assessment by students in addition to evaluation by others. The intent is to find out about a particular student in a particular setting over a period of time, and giving him the opportunity to show what he does best in the way he does it best. Undoubtedly, there are numerous difficulties associated with such a paradigm: among them are that teachers and students need far more power than they have at the moment and they need to be very clear about the nature of performance assessment, schools, curricula and schedules have to be

restructured to accommodate more flexible assessment practices, and society has to stop evaluating schools in terms of statistical scores. Despite the problems, however, personalized alternative assessment may at the moment be the only way to counter the absence of student-centeredness in standardized testing.

Rubrics and scoring

Assessment reform involves more than merely changing the nature of the tasks. A lot of thought has to be given to what kind of corresponding change there should be in the way performance tasks are scored. For example, some teachers vehemently oppose giving grades, so should grades be used at all? If they are not, how does one document student achievement? Keeping descriptive records of students' progress and performance is one option. Portfolios of students' work is another. But if grades have to be given, what are the standards by which they are determined? There is no easy answer. For example, although the Vermont portfolio assessment program was an attempt to move away from standardized testing, their writing assessment still used an assessment rubric that required every student to meet the same specified criteria broken down analytically into dimensions such as purpose, details, organization and mechanics. Essentially, the students were still being evaluated in a standardized manner.

Test security

The perceived need for test security is another major difference between standardized testing and alternative assessment. The assumption that there are certain `correct' answers necessitates keeping the content of tests confidential, while assessment calls for open negotiation between assessorand student.

One reasonable argument for supporting test security is Wiggins': open knowledge of test content may cause students to focus only on limited domains of learning. However, open performance-based assessment can be designed to ensure the need to integrate knowledge from various domains. Test situations that I can think of which may warrant security are ones which need to assess the appropriateness of spontaneous response, such as spontaneous medical decisions; in such cases, I feel it is justifiable to keep both test content and format unknown to the student before the test, but discussion on results must be open.

There are, however, many reasons why tests should *not* be secure. The first concern is with authenticity. Many forms of assessment in the real world beyond educational settings do not practice the kind of secrecy adhered to in school settings, particularly in standardized testing; thus, going through secure tests is not an authentic assessment experience. Security generally requires one-shot testing in

artificially constructed settings (for example, secure rooms); subsequently, the validity of interpretations made of students' achievement in secure tests can be questioned, since it is not an assessment of authentic use of knowledge.

Secondly, although test security is intended to ensure credibility and fairness in test administration, the irony is that secrecy actually leads us to challenge those very aspects. Secrecy cuts off dialogue between student and assessor and makes it impossible for students (as well as teachers and parents, in the case of state-wide tests) to raise objections or ask for clarification before tests are administered. Inevitably, this leads to doubts about the credibility of the tests, although objections are rarely entertained. In some situations, hours- long quarantine may be imposed on students before an exam in which there is suspected leakage of test information. Both the validity and reliability of student performance under such conditions can be seriously called into question; more often than not, however, no one does so, particularly not the powerless students, even though they are the real victims. Thus, the moral dimension of secrecy in testing is also a subtantial cause for concern. In addition, to students who are denied knowledge of the standards and criteria by which they are judged, tests are both arbitrary and unfair. Secrecy in high-stake testing also leads to high anxiety, which in turn often results in unethical practices, such as selling and buying 'leaked' test questions, and cheating during exams.

Thirdly, there can be very little or no instructional value when results are kept secret even after the test. Students are also denied consultation with or guidance from adults during assessment, nor do they have access to resource materials, restrictions that do not often exist in the real world. In addition, because the whole testing procedure is shrouded in secrecy and only selected personnel are allowed access to test planning and design (in standardized tests), test security leads to an over-reliance on and excessive belief in 'expert' assessment authorities and a distrust of teacher competence. Thus teachers have little control over testing imposed by strangers on their students, whom they know best.

Reliability and validity

The distinctions between standardized testing and alternative assessment are tied in to issues of reliability and validity. Both Wiggins' and Berlak's texts for this topic address these constructs to some extent. *Reliability* refers to the consistency or replicability of scores or performances, while *validity*, as Messick defines it, is the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989). Both standardized testing and alternative assessment attempt to achieve a degree of reliability and validity. Standardized tests, however, have a greater concern with reliability and are "intrinsically prone to sacrifice validity"

(Wiggins, p.4); they are usually simplified for the sake of precision in scoring, and thus lose the authentic complexity of real-life tasks. In addition, the test security demands test-taking under artificial circumstances, as has been previously mentioned, rendering the validity of the inferences one can draw from test scores produced under such circumstances questionable. Alternative assessment, on the other hand, by focusing on more authentic ill-structured tasks, considering individual students to a greater degree, and allowing negotiation between the student and teacher, reduces the degree of reliability but raises the degree of validity of the inferences that can be made about the student's performance. The Hot Topic guide on Measurement Issues discusses these measurement constructs in greater depth.

Conclusion

Like most other problems that impact society in a significant way, the issues surrounding educational testing and assessment are not easily resolved because they are tied to ideology, politics and social systems. Standardized testing is familiar and comfortable, while numerous obstacles still have to be overcome in order to successfully reform assessment. Standardized test results are the basis for reproducing a stratified society, but true assessment reform, particularly assessment without traditional grades, may lead to the restructuring of society. The question is whether society is ready to move into a new paradigm.

References

(not listed in this Hot Topic guide)

Books:

Mabry, L. (1992) Performance assessment. In Debra D. Bragg (Ed.). Alternative approaches to outcomes assessment (p.109-128). University of California, Berkeley: National Center for Research on Vocational Education.

Messick, S. (1989) Chap. 1 in Robert L. Linn (Ed.) Educational Measurement. NY: American Council on Education, Macmillan.

Stenmark, J.K. (1991). Mathematics assessment: Myths, models, good questions and practical suggestions. Reston, VA: NCTM.

Articles:

Archbald, D.A. & F. Newmann (1992). Approaches to assessing academic achievement. In Harold Berlak, Fred M. Newmann, Elizabeth Adams, Doug A. Archbald, Tyrrell Burgess, John Rave & Thomas A. Romberg Toward a new science of educational testing and assessment. Albany: State University of New York Press.

Newmann, F. and D.A. Archbald (1992). The nature of authentic academic achievement. In Harold Berlak, Fred M. Newmann, Elizabeth Adams, Doug A. Archbald, Tyrrell Burgess, John Rave & Thomas A. Romberg Toward a new science of educational testing and assessment. Albany: State University of New York Press.

Shepard, L.A. & M.L. Smith (1989). Escalating academic demand in kindergarten: Counterproductive policies. In The Elementary School Journal Vol. 89, No. 2. p.135-45.

Chap-1

From: Harold Berlak, Fred M. Newmann, Elizabeth Adams, Doug A. Archbald, Tyrrell Burgess, John Raven, & Thomas A. Romberg (1992)

Toward A New Science of Educational Testing and Assessment.

N. York: State Univ. of New York Press.

1

The Need for a New Science of Assessment

Harold Berlak

Introduction

The idea that schooling for all is essential for social progress and economic growth grew up alongside the development of industrial capitalism during the tail end of the nineteenth and early decades of the twentieth century. By the 1990s, the aspiration for universal schooling has come a long way toward realization, though many American youth still do not complete secondary school.[1] While universal provision of schooling is still widely seen as a noble, if unrealized goal, there is a growing consensus that the system of public education that has evolved over the course of this century in the United States is in serious trouble. Public officials, corporate leaders, and ordinary citizens are increasingly dissatisfied with the quality of the education provided by the nation's schools to the great majority of children. While the margins of the American political scene, left and right have long been critical of schools (albeit with quite different ideas of the problems and solutions), with the exception of racial desegregation, discussions of elementary and secondary schooling policy over the last 25 years were virtually absent in the national media, in the platforms of the national political parties, or in campaigns for national state or even local public office. For brief interludes, following the launching of Sputnik in the late 1950s and in the mid-1960s during Lyndon Johnson's "war on poverty," public attention focused on schools, but this interest was not sustained.

This changed in 1983 with publication of *A Nation at Risk*, a report of the National Commission on Excellence in Education (1983). It made national news with its assertion that American education was threatened by "a rising tide of mediocrity," and with its frequently cited lines: "If an

unfriendly foreign power attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. As it stands, we have allowed this to happen to ourselves. . . . We have, in effect been committing an act of unthinking unilateral disarmament."

Why this report received so much attention is a matter of some conjecture. Very serious problems, particularly in, but not restricted to, inner city and poor rural schools, had existed and been widely known for many years. In spite of the report's claims to the contrary, what had changed were not the problems[2]—though undoubtedly they had gotten worse—but the public's and elected officials' response. The reason for wide notice of *A Nation at Risk* had more to do with the particular historical moment it appeared than with the originality or profundity of its analysis. In the early eighties, the failures of the US economy had just begun to penetrate the nation's consciousness—dominating the news were the galloping US trade deficit; the failures of US industry; plant closings; and dramatic increases in unemployment, particularly in the older industrial cities. What this report offered was an explanation for these apparently inexplicable events, an explanation which was eagerly embraced by the mainstream press and corporate America, and widely repeated in the national media. The report told the American public that a major cause, if not the major cause, of America's fall from grace as the world's preeminent economic and industrial power was the failure of the nation's schools to educate a competent, dedicated work force. This was a palatable diagnosis of the nation's economic malaise that suited the times. It placed blame, not on the basic structural problems of the US economy, nor on the failures of corporate leaders and politicians to address the changing world economy, and to do something to relieve the accumulating social problems and the gross disparities between rich and poor; but on the politically impotent: the nation's elementary and secondary school teachers, nameless educational bureaucrats, and unskilled and/or unmotivated workers.

A Nation at Risk was not the work of right-wing ideologues. Terrell Bell, who initiated the report, and who was appointed by Ronald Reagan as his first secretary of education, was at the time widely regarded as a middle-of-the-road professional, and the eighteen-member National Commission on Excellence Bell appointed included, among others, the retired chairman of the board of Bell Laboratories, two professors from Harvard and University of California at Berkeley respectively, four university presidents (including Yale), a former governor of Minnesota, the immediate past-president of the National School Boards Association, two principals, two school board members, the superintendent of schools from Albuquerque, and the 1981–82 teacher-of-the year, a high school foreign language teacher from an affluent suburb of New York City.

Whatever its deficiencies, the *Nation at Risk* drew public attention to the schools, and this, attention contrary to the expectations of many, has continued to the present. The report and the wide attention if received stimulated responses from virtually every organization and group with an interest in educational policy. Since 1983 countless reports, articles, and books have been written or commissioned by every major foundation, dozens of minor ones, policy think-tanks across the political spectrum, associations of corporate executives and educational professionals, teachers' unions, children's and parents' advocacy groups, formal and *ad hoc* organizations of state and local educational officials, as well as by individual journalists and scholars. While there are major differences in the policy recommendations, very few reports contest the *Nation at Risk's* view of the economy, and none with dissenting views have received wide public notice.[3]

All this talk about education did, however, galvanize latent public discontent with the schools and create a political climate for change. Since 1983 virtually every governmental agency and administrative unit at the state, county, and school district levels that held some responsibility for elementary and secondary schools has initiated and implemented some reforms. State legislatures, governors, state and local education officers, the major foundations and think tanks, the two leading national teachers unions, and even the 1988 presidential candidates, Bush and Dukakis, felt the need to respond to the clamor for educational excellence.

Many of the responses can be passed off as media hype and political rhetoric. But there were also many concrete measures undertaken. I make no effort here to recount and analyze these efforts in any detail, a monumental undertaking far beyond the purview of this chapter. However, some effort to make sense of these intended reforms is essential if we are to understand the current movement for developing new forms of educational assessment and testing.

An Analysis of the Reform Movement: The Role of Testing

Two competing tendencies about how political decisions should be made and who should make them are represented by recent efforts to reform the nation's schools. One tendency is toward decentralization of authority and decision-making by those who are most immediately affected by those decisions. This view is often coupled with a distrust of centralized authority and a disdain for experts and intellectuals. From this perspective, "bottom-up" change is valorized along with direct, grassroots or participatory democracy.

The second tendency in this society is toward centralization of authority and decision-making, with responsibility for the difficult decisions left to the

man or woman at the top—the CEO, the chief of staff. In the case of schools, the superintendent or principal must be a tough-minded leader, able to shape up the troops, delegate responsibility and hold subordinates accountable for their performance. Efficiency and immediate, demonstrable results are valorized, and while democracy is not necessarily rejected, it is representative democracy and delegation of authority to those who know best which is endorsed—with little tolerance for participatory democracy, which is seen as chaotic and in the end as encouraging the lowest common denominator in terms of process and product.

The relative strength of these two tendencies and the ambivalence many Americans feel about how to reform schools are evident in the multiplicity of proposals advanced and policies instituted since 1983. The language that has dominated the discourse about school reform has been that of crisis, of disaster, of imminent threat to the very survival of the nation. I have already quoted *A Nation at Risk* with its military metaphors. Here are the words of *A Nation Prepared*, the second-most influential report, published by the Carnegie Forum on Education and the Economy (1986), created and supported by the Carnegie Corporation of New York:

> American's ability to compete in the world markets is eroding. The productivity growth of our competitors outdistances our own. As jobs requiring little skills are automated or go offshore and demand increases for the highly skilled, the pool of educated and skilled people grows smaller and the backwater of the unemployable rises. Large numbers of American children are in limbo—ignorant of the past and unprepared for the future. Many are dropping out—not just out of school but out of productive society.
>
> As in past economic and social crises, Americans turn to education. They rightly demand an improved supply of young people with the knowledge, the spirit, the stamina and the skills to make the nation once again fully competitive. (p.2)

In times of national crisis, it is no surprise that the strongest impulse by politicians most directly responsible for schools is to use their authority by employing the tools they understand and know best. In the United States, basic responsibility for schools resides with the states. Eight years after publication of *A Nation at Risk* virtually every state had instituted a combination of top-down measures intended to raise educational standards. These measures include requirements for academic courses, new or strengthened controls over textbook adoptions, mandated use of state curriculum guidelines which in some instances are closely aligned to required tests, and more pre-

scriptive regulations for certifying teachers. But, by far the most common measure is statewide testing programs throughout the grades that, in effect, increased the proportion of education dollars spent at the state level, and strengthened the control of the state's chief educational officer and/or state department of education.

While it is difficult to generalize about several thousand school districts, many, particularly the larger urban systems, responded much like state departments of education by tightening and centralizing bureaucratic control over curriculum, pedagogy, grading, student discipline, and personnel selection. In addition to the newly devised or revised state "basic skills" tests, and the standardized achievement tests which have been used for many years almost universally throughout the grades, some districts instituted their own district-wide tests, in some cases going so far as to specify textbooks for each grade level, and to link mandated tests to these texts.

The role of the federal government under Reagan-Bush is contradictory. On the one hand their administrations greatly reduced or eliminated programs supporting educational research and development, curriculum and staff development, as well as programs that aided particularly needy populations, using the justification that schools are primarily the responsibility of local and state governments. On the other hand, the Department of Education, whose elevation to cabinet-level status was bitterly opposed by Reagan and right-wing groups prior to 1980, in the ensuing years became an increasingly active instrument in efforts of right-wing forces within the federal government to shape local and state schooling policy through, for example, selective enforcement of and in some cases opposition to agreements reached by local and state school officials and the courts on civil rights issues, active advocacy of a national core curriculum, national assessment, and so-called "freedom of choice" plans which would, in effect, divert public funds to private schools. Among the more visible efforts by the federal government to shape schooling practice is the annual media event staged by the secretary of education upon publication of the "wall chart," which ranks the states' educational performance based on standardized test scores. In some instances a form of this annual ritual is repeated by states publicizing rankings of school districts, and by the central administrations of school districts releasing to the press rankings of individual schools within districts.

What explains the enormous emphasis on tests? I have suggested that a primary reason for this emphasis is that tests are a means of maintaining centralized control, providing those higher up in the educational bureaucracy (central office administrators, school board members, state education officials, legislators, etc.) with relative rankings of organizational units (classrooms, schools, districts, etc.) and/or students and teachers. This, however, is not an adequate explanation since it does not account for widespread popular

support for the use of tests. While there is increasingly vocal criticism of tests among professionals and by the national media, there is still remarkably little evidence of widespread discontent with current forms of testing. Indeed, many support increased testing, including African-American, and Latino-American parents who are convinced that their children, who consistently score lower on standardized and criterion-referenced tests, have been and continue to be victimized by low expectations on the part of teachers and school officials. For many within these communities, the only credible indicator of improved educational performance is improved performance on standardized tests. The irony in this is that, while the demand for more professional accountability is certainly justified, any gains on such tests are often temporary and local. The technology of these tests assumes there will be winners and losers, and in our society the winners are invariably the more affluent and the losers the poor and powerless.

Efforts to reform schools from the center continue, but a counter tendency toward more democratic school-level control has become more visible recently for several reasons, including organized opposition to centralized control by teachers unions, parent groups, and local school boards, and a growing conviction that mandating changes from above has not worked. What a few years ago was a fringe view that genuine changes in the end must occur in individual classrooms, which is not possible without active participation of teachers and without a large measure of autonomy within each school, has become increasingly accepted as the common wisdom by the public policy establishment and the mainstream press.[4]

Several states while tightening centralized control, have encouraged school–level decision–making by altering state regulations to permit principals and teachers more say about school expenditures, curriculum and staffing. Also several districts scattered across the country—New York City, Buffalo, and Dade County, Florida, are the most frequently mentioned in the press—not only tolerate but appear to foster school-level decision-making. However, although talk about, and arguments for, teacher empowerment and school-level governance are commonplace, it is the rare exception rather than the rule for central office bureaucracies to yield power.

This ambivalence over who should call the shots, the authorities at the center or the local school community, is probably nowhere more clearly exemplified than in the previously cited Carnegie report, *A Nation Prepared*. On the one hand, the report celebrates the role of the teacher and provides what it calls "a scenario," a hypothetical example of a high school run by the school staff in close collaboration with the local community. On the other hand, however, the report makes no recommendations as to how centralized administrative control by school districts or the state is to be relinquished. Its

key and sole concrete proposal is creating a new National Board for Professional Teaching Standards which would, in effect, centralize the certification of an elite cadre of master or lead teachers whom they assume would transform the schools.

If there is any consensus after almost eight years of intensive public discussion and activity, it is that tinkering with regulations and issuing more administrative mandates will not suffice, and that what is needed is *perestroika*, a basic restructuring of the entire system. *Restructuring* is one of those words like *democracy* and *accountability* that have an inexhaustible number of possible meanings, each aflame with ideological passion. At very least it implies an unfreezing of the central office bureaucracy and a shift in authority and the power of decision-making from existing to new formations.

In spite of the calls for *perestroika*, decentralizing authority, and empowering teachers and principals to institute changes from below, there has not been any wide-scale restructuring of the system. Except for some well-publicized exceptions, the evidence is that, overall, the system has become more and not less centralized over the past eight or so years. (Sarason, 1989) While there are several interconnected factors at work, one—if not *the*—single most significant in holding the current system in place, indeed in strengthening the current structures, is testing. Not any tests, but the *particular forms* of standardized and criterion-referenced testing which have become the main instruments of reform. Here we have the major paradox of the reform movement of the eighties: significant improvements in the quality of schooling are impossible without structural changes, but increased dependence on mass-administered tests at all levels has had the effect of strengthening existing structures and forms of control. The culprit is not educational assessment and testing *per se*. Rather, the argument I make here and in Chapter 8 is that the particular forms of testing in widest use for increasing accountability are rooted in a social science paradigm which takes as a given the necessity for centralized control.

Use of such tests are not the sole cause for the failures to restructure schools. *Re*-forming schools or any social institution is a complex business. It requires a commitment by national, state, and local, public officials, and professional educators to critically examine their own long standing practices and patterns of organizational control. It takes persistence and inordinate courage by leaders and governing bodies to dislodge entrenched, centralized bureaucratic power. If we know anything at all about politics and human behavior, it is that many endorse the need for change, but few risk challenging the many vested individual and institutional interests in maintaining business-as-usual. There are thousands of organizational entities, and tens of thousands of individuals within national and state governments, colleges and universities,

foundations, publishing companies, and central offices of local school districts whose power would be greatly diluted or lost if the current system of assessment were significantly altered.

The historically unparalleled growth in the use of mass testing as the chief instrument of school reform over the last several years has produced a counter-reaction as evidenced by increasing public criticism in mainstream journals and the popular national press questioning the credibility of these tests, and by a resurgence of interest in alternative forms of assessment. Two recent studies, the first conducted by the National Center for Fair and Open Testing (Medina & Neil, 1988) and the second by the National Commission on Testing and Public Policy (1990) document both the growth of and interest in the development of alternative forms of testing, and the resistance to use of current forms of testing by many mainline educators and citizen and professional groups. Skepticism of multiple choice tests, which for many years was largely confined to progressive critics and to academic traditionalists, is now voiced regularly in such places as the *Washington Post*, *New York Times*, *Wall Street Journal*, *Newsweek*, and even on prime time television documentaries.

The two reports cited above and a publication of the National Center on Effective Secondary Schools at the University of Wisconsin (1989) document in detail the deficiencies and problems with these tests. They show that the short-answer, closed-ended format precludes the assessment of higher-order thinking and mastery of complex material, that test items are frequently biased in subtle and not so subtle ways, and that dependence on these tests as the primary indicators of school quality and for making judgments about students abilities and achievements distorts schooling policies and practice in numerous ways.[3]

Though I (and all the writers included in this volume) would concur with most of these criticisms of the commonly used forms of educational tests, and that there is a need to develop alternatives, I do not focus here on critique nor on reviewing and examining proposed alternative forms of testing. Rather my purpose in this chapter is to raise questions about the theoretical foundations of the widely used forms of achievement testing, and to foreshadow the argument for a theory of testing and assessment, that is compatible with current interest in restructuring schools by dispersing power and shifting responsibility away from the center, towards local school districts and to the teachers and principals within individual schools.

Though there is critique in this volume, and discussions and exemplars of alternative forms of assessment, the book is primarily an effort to examine the theory and practice of educational assessment, and a modest step toward the development of a new paradigm. This book supports the view that fundamental changes in the way we think about education and the process of schooling must accompany the effort to rethink assessment theory and practices if we are to realize the aspiration of providing all the nation's children with schools which serve their best interests, the interests of the communities they live in, and the interests of the nation as a whole.

I must forewarn the reader that this book does not pretend to provide a fully articulated and coherent perspective on the theory and practice of educational assessment. The lack of unity and consistency of argument across chapters is, in part, a function of its history. Supported by a grant from the US Department of Education's Office of Educational Research and Improvement to the National Center on Secondary Education at the University of Wisconsin, I collected and edited a set of papers which were intended to provide some fresh perspectives on the testing and assessment question drawing upon work commissioned by the Center and from the existing assessment literature. This task was completed in 1988. In the course of this work, it became increasingly clear to me that some of the researchers whose writings I had collected and edited were pressing the limits of the familiar testing technology and moving in the direction of abandoning and replacing the measurement paradigm which has predominated for at least the last sixty years. Five chapters in this book are revised and edited versions of papers selected from that earlier collection, and three chapters (Chapters 1, 6, & 8) were written expressly for this volume. The first and last chapters are an effort to illuminate the arguments for a new assessment paradigm, arguments which I saw as largely submerged in the work of the writers of the other papers. In none of the chapters, except Chapter 5 by John Raven, and my two chapters, is there a self-conscious effort to articulate a case for a new science of testing and assessment. Although I make my case drawing freely from the work of others, from the writers of the other chapters, and from sources I cite in the endnotes of my two chapters, I alone must be held responsible for the way I have interpreted and used their work.

Foundational Assumptions of the Current Paradigm

I will state what I see as the four foundational assumptions of the paradigm which underlies virtually all standardized and most criterion-referenced tests. In so doing, I will also state four "counter assumptions" which are intended to foreshadow the argument for the development of a new testing and assessment paradigm.

Before proceeding I will clarify several commonly used terms:

Test Technology. Test technology refers to the structure of a test, the ground rules and conventions used for its construction, the procedures and protocols for scoring and summarizing results, and the matrix of practices required for everyday use.

The tests I refer to here are those generally composed of a relatively large array of short questions of "items." Each item includes a problem presentation—a sentence, paragraph, set of statements, a chart, graph, picture, or mathematical equation followed by a set of four or five possible responses, one of which is designated by the test-makers as the correct or best possible answer. The individual taking the test makes a selection and blackens a space provided, generally on an separate answer sheet which is subsequently machine scored. There is almost always a time limit for completing the test. Scores are usually computed by counting correct responses and subtracting this number from the number of incorrect responses. A variety of statistical operations is employed for summarizing test results so that they may be used for comparing scores of individual or groups. Some variations of this technology should be noted, which generally do not represent a significant change in a test's technology. A desktop computer or terminal may be used to present items to the test-taker and to tally responses in lieu of the printed test and answer sheet. Also, some tests may include open-ended test items, those which require a writing sample or solving a math problem. In scoring such items, responses are assigned a number by a person trained in the use of a set of scoring conventions. The scores are then treated in the same way as those derived from multiple choice items.

Standardized and Criterion-Referenced Tests. A distinction is commonly drawn between "standardized" (or norm-referenced) and "criterion-referenced" tests. Among the best known of the former are the California Achievement Tests, the Iowa Tests of Basic Skills, and the Standard Achievement Tests (or SAT). Criterion-referenced tests include virtually all National Assessment of Educational Progress (NAEP) tests and state-mandated "basic" or "essential" skills tests.

Standardized tests do not depend upon setting educational standards as is often assumed. The concept of standardization in this context refers to tests which are constructed in such a way that allows a standard score, grade equivalency, or percentile to be computed, thereby permitting comparison of an individual's score, percentile, or a group mean to those of another individual or group. Such comparisons are possible only if the test is "normed." What this requires is that during a test's development, it was administered to a sample of test-takers, and the distribution of their scores was compared statistically to a so-called "normal" distribution. The slope of such a distribution is bell-shaped, hence the commonly used term *bell curve*. A normal or bell curve does not appear naturally. To the contrary, test-makers attempt to compose test items so that there will be a suitable ratio of correct to incorrect responses. If too large a number of test-takers chooses the correct responses to sets of items, these items would be revised or abandoned even if there were unanimous consensus that the items tapped an educationally significant body

of knowledge or set of skills. The reason is that the items must "discriminate," that is, produce the proportion of correct to incorrect answers required by a "normal" distribution.[6] The technology of standardized tests, contrary to popular belief, do not warrant making *qualitative* statements about a person's (or group's) performance. The only claims which are warranted is how an individual's score or percentile (or group's mean or mean percentile) compares with others who have taken a version of the same test.

Though there are a number of recent efforts by the NAEP and several states to depart from the usual closed-ended format, the items in the vast majority of criterion-referenced tests are indistinguishable from those included in a standardized test. The major difference is that criterion-referenced tests are not normed. A panel of educators decides what percentage of correct responses constitutes passing or minimal competence. This score serves as the criterion for making judgments about an individual's or groups' competence or level of achievement. In practice, someone selects a score which sets the minimum number of items students at a particular grade level must answer correctly in order to be considered minimally competent in a given area—mathematics, reading, or whatever. Criterion-referenced tests (with some significant exceptions) also warrant only quantitative statements about how an individual's score or a group's mean (the group may be a single class, a school, a set of schools from a district or entire state or region) compares to the mean of another individual or group, or to an established criterion score.

It is important to note that in recent years, there have been efforts to develop so-called "performance-based" tests. The intent is to create assessments which avoid the multiple choice format and more closely approximate real tasks, such as conducting an experiment or writing a job application letter. While some of these efforts succeed in breaking the boundaries of the conventional testing paradigm, most do not depart significantly from the conventional standardized and criterion-referenced test technology. Rather than presenting four or five alternatives to choose from, a score is assigned to the test-takers' "free" responses (recorded on paper or computer) on the basis of previously-determined criteria. Aggregate scores are then treated in more or less the same way as those derived from multiple choice items. For all practical purposes most such assessments are rooted in the conventional psychometric paradigm.

Scientific Paradigms. Scientific endeavor in any area rests upon a set of *a priori* assumptions shared by persons who engage in that endeavor. With reference to testing, this means that those within the educational testing and evaluation community who design and construct educational tests, or who administer and interpret their meaning to others take for granted a set of beliefs, values, and practices. (or "puzzle solutions"). It is the foundational assumptions and practices taken as normal within a particular community of

scientists which Thomas Kuhn, a well-known historian of science, calls a *paradigm* in a widely quoted book, *The Structure of Scientific Revolutions*, first published almost thirty years ago. A paradigm may be seen as what Michael Foucault calls a "regime of truth." A regime of truth in science is a set of practices and discourses taken as given in everyday scientific activity and which implicitly defines what are and are not considered legitimate scientific questions and methods.

What is significant to my argument here, and to the thesis of this entire volume is Kuhn's claim that paradigms or regimes of truth in science are transient and that the history of science is itself a history of paradigm breakdown and replacement. Paradigms are replaced because anomalies and problems appear that cannot be explained or be fruitfully addressed using the commonly accepted language, ground rules, or "puzzle solutions." Over time scientists develop new paradigms—that is different concepts, sets of "puzzle-solutions," and a constellation of beliefs and values[7] which appear to address the difficulties. It is these changes that constitute revolutions in scientific thinking and practice, and while they are infrequent, major transformations are to be expected sooner or later. In the meantime, normal science continues more or less undisturbed, as the old regime erodes and in time is replaced by a new one. Periods of transition and change, it should be added, are unsettling if not tumultuous because the new paradigm threatens existing interests and the institutional arrangements that hold the current regime of truth in place.

The scientific paradigm that undergirds standardized and virtually all criterion-referenced tests which has been in the process of breakdown for the last two decades has reached a critical stage. Standardized and criterion referenced tests, rooted in an anachronistic paradigm, are a major barrier to the renewal and restructuring of the nation's schools. As we enter the last decade of the twentieth century, it is becoming apparent, at least to those outside the testing and measurement establishment, that the assumptions which are intrinsic to the technology of standardized and most criterion-referenced tests are untenable. Out of the ashes of this paradigm, from the many varied and imperfect efforts underway to solve the practical problems of assessing educational achievement, is slowly emerging a new paradigm, one based on a set of foundational assumptions that are in sharp contrast to those that underlay the current paradigm.

The paradigm that is foundational to current forms of standardized and criterion-referenced tests I label the *psychometric* paradigm; the emerging one, a *contextual* paradigm. There is some risk in the use of these terms, as there is in any effort to classify and simplify complex ongoing human activities into categories. The distinction is helpful insofar as it helps to clarify the issues and distinguish significant differences in efforts to develop alternatives to the most commonly used forms of testing and assessment. The implication

that all tests and assessments may be classified in terms of two mutually exclusive categories, however, is potentially misleading and confusing because the distinction also may obscure significant differences within and similarities across categories. As Doug Archbald's and Fred Newmann's summaries of alternative forms of assessment show (see Chapter 7), some efforts appear to embody aspects of both paradigms.

It should also be underscored that the psychometric paradigm must not be considered as synonymous with *quantitative* methods, and the contextual paradigm with *qualitative* approaches. It is certainly true that psychometric assessments rely heavily on quantification and statistics, and contextual assessments more often than not employ qualitative methods. However, quantitative measurement and the use of statistics are not necessarily inconsistent with contextual approaches, and qualitative techniques are sometimes used in ways that ignore or bypass social context.

Assumption 1: Universality of Meaning. By universality I mean a view that there is or can be established a single consensual meaning about what standardized or criterion-referenced tests claim to measure which, in effect, transcends social context and history. For example, a standardized reading test purportedly indicates a person's ability to read in the real world, not just in the testing situation, and "ability to read," it is assumed, has a more or less universally understood and accepted meaning. Further, it is assumed that scores on a given reading test indicate individuals' level of reading ability—regardless of their or their families' history, culture, or race; regardless of gender, whether they live in Nome, Alaska or Newark, New Jersey; regardless of whether they have gone to a school with a first-class library or no library at all, and whether they reside in an affluent suburb or an area with high and chronic unemployment. The assumption of universality, in effect says, that a reading test score has essentially the same meaning for all individuals everywhere.

Within the discourse of psychometrics, postulated attributes or capacities of persons (their reading ability, or academic achievement in a particular area, for example) are called *constructs*. A standardized test of academic achievement presumably measures the *construct* of "academic achievement"; a criterion-referenced test of basic or essential skills measures the *construct* of "basic" or "essential skills." The term *construct* may sound strange and may perhaps be considered superfluous to non-specialists in the field of educational measurement. This term became commonplace in the field of mental measurement after its use in a seminal article by Lee Cronbach and P. E. Meehl titled "Construct Validity in Psychological Tests" (1955). According to Cronbach and Meehl a construct "is an intellectual device by means of which one construes events. . . . It is a means of organizing experience into

categories. . . . Construct validity, then, is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (pp. 281–82). The use of this term acknowledges an obvious but sometimes ignored fact that human attributes or capacities are not tangible, directly observable, or measurable. Thus, a reading test does not, indeed cannot measure reading directly. Rather, if the reading test does what it claims, it measures a construct the test-makers have labeled "reading" or "reading ability."

How do we know whether a test measures what it claims to measure, whether a test in fact measures authentic reading ability or genuine academic achievement? The response a traditional testing expert gives to the question of whether a standardized or criterion-referenced test measures what it purports to measure is that this determination depends upon the adequacy of the case a test-maker makes for the test's *construct validity*.

Establishing construct validity of a test requires getting things straight between (1) the world of human events and experience, (2) the construct label, and (3) the test. This entails establishing what Cronbach and Meehl refer to as a "nomological net," which is "a rigorous (though perhaps probabilistic) chain of inference" from an empirical body of knowledge and a logical analysis of the meaning of the construct. Almost thirty years later, Cronbach (1987) stressed that "the argument [for test validation] must link concepts, evidence, *social and political consequences and values*" [italics added]. Thus, in order to establish the validity of a test of academic achievement within the framework of the psychometric paradigm, for example, one would need to assume that the construct of "academic achievement" has a stable, universal meaning, or that unanimity on its meaning is both possible and desirable, *and* that it is possible to reach consensus on the desirability of the social and political consequences of the test's use.

The counter assumption: plural, and contradictory meanings. In Chapter 8, I examine in some detail the basic controversies and contradictions in contemporary America over education and the functions, purposes and practices of schooling. I demonstrate that the assumption that there can be a meaningful nationwide statewide, district-wide, or even schoolwide consensus on the goals of schooling and on what students should learn and how they should learn is untenable. I also argue that in a muticultural society which values difference, consensus is undesirable. While it is perhaps understandable that in the 1950s some would hold the view that consensus on basic educational beliefs and values is possible, from the vantage point of the 1990s this view is naive. The premise of what I call a contextual paradigm is that a plurality of meanings, and differences and contradiction in perspectives are inevitable in a multicultural world, where individuals, and groups have differing histo-

ries, divergent interests and concerns. There is no, nor can there be universal consensus on what constitutes "ability to read", the meaning of "academic competence" or "authentic achievement" in general terms or within specific academic fields. Experts and nonexperts alike hold plural and often fundamentally contradictory beliefs and values over the meaning of all educational terms. Validating educational tests based on psychometric canons represents a quest for certainty and consensus where certainty is impossible, and agreement is unlikely unless differences are suppressed and consensus is overtly or covertly imposed. Further, as I argue in Chapter 8, the entire concept of "construct validity" on which the scientific credibility of all such tests is rooted is itself internally contradictory and untenable. I also argue that it is possible to develop a system of educational assessment that takes plurality of perspectives and differences in values and beliefs as givens, and treats these differences as assets, rather than obstructions to be overcome.

Assumption 2: The Separability of Ends and Means, and the Moral Neutrality of Technique. Discourse and practice within psychometrics assume that tests, if constructed and interpreted according to accepted standards, are *scientific instruments*, which are value-neutral and capable of being judged solely on their scientific merits. The argument often made in defense of the technology of standardized and criterion-referenced tests is that their development represents an advance over prescientific and subjective forms of assessment, such as grades and teacher-made tests, which intermingle factual observations with the personal, subjective dispositions of the teacher. The basis of the argument for the moral neutrality of tests is that ends and means are separable. Questions such as what constitutes the good or just society, or what is the nature of a good or proper education, because they require moral choices, are not resolvable, and hence lie outside the domain of true science. The choice of means or the best route to a prescribed goal or end, however, is seen as an empirical matter, not a moral question, and hence may be decided scientifically. From this perspective, the job of the assessment expert parallels that of the engineer whose expertise is in the application of the science, not in making judgments about desirability or worth of the enterprise. The role testing expert is limited to dealing with technical or procedural questions within the moral framework set by society.

There are two closely connected assumptions here. First, facts and values, (or what is and what ought to be) are distinct and separable, or they are sufficiently distinct to make possible a non-normative science of educational measurement. What follows from this assumption is that testing experts can make technical decisions without making value judgments. Second, the assessment scientist is best equipped to make judgments about means, that is to develop the ways of assessing educational outcomes and how these are to be

properly used and interpreted. Just as it would be the height of irrationality to turn over to a non-engineer the responsibility for designing a bridge or a rocket's guidance system, so too would it be irrational to replace scientific techniques of measurement and the rules of evidence with the opinions and subjective preferences of the non-scientist.

Counterassumption: The Inseparability of Means and Ends. The impossibility of sustaining this fact-value distinction is argued in Chapter 8. In brief, the argument for whether a test measures what it claims to measure rests on the case made for its construct validity, which is considered a technical matter. However, establishing construct validity clearly is not merely a matter of empirics, getting the facts straight and interpreting them according to established rules of evidence. Judgments about an educational test's validity invariably require choices among contradictory values, beliefs and schooling practices (Cherryholmes, 1989; Messick, 1989). In the real world of schooling, separating means and ends is not possible. All assessment procedures have the power to directly or indirectly shape social relationships—how students, teachers, and administrators within a setting interact with one another, what they will or will not say or do in particular situations. Moral questions arise in all social relationships, which can either be resolved by the use of direct or indirect power where the values, beliefs and ideologies of those with the ability to impose their will prevail, or by a process wherein conflicts are acknowledged, and mediated recognizing both differences and commonalties in interests and values. If judgments about assessment procedures and testing are left to experts, then they assume the responsibility for resolving differences over basic moral questions which in a democratic society should be settled by ordinary citizens and/or their democratically elected representatives.

Assumption 3: The Separability of Cognitive from Affective Learning. The psychometric paradigm separates the assessment of learning outcomes and processes into distinct and mutually exclusive categories, separating cognition or academic learning from affect, interests, or attitudes. Sometimes a third category, psychomotor outcomes, is added to the set. Tests of academic achievement, and IQ tests fall into the first category; tests or inventories which solicit a person's beliefs, attitudes, or interests fall into the second; and tests of a person's capacity to perform a hands-on or vocational task (such as auto mechanics or typing) fall into the third. This three-way classification of human learning or capacities divides head, heart, and hand, that is, it separately assesses those areas of human learning and development related to the realm of the intellect, those related to the realm of feelings and values, and those which require manual or physical dexterity. A test of basic educational skills, for instance, purportedly will tell us how well a person knows a particular body of scientific facts or performs a particular set of math tasks. If we want to know the person's interest in math, or whether she is curious about science, we would need to administer a different instrument.

These distinctions are deeply ingrained and institutionalized within the psychometric sciences and are rarely given a second thought. They are legitimated by the Benjamin Bloom's (1956) *Taxonomy of Educational Objectives* which remains the most widely accepted system of classification in the field of education. The distinctions are treated as virtually self-evident and used widely in the everyday discourse of teachers and administrators.

The Counterassumption: The Inseparability of Cognitive, Affective and Conative Learning. As John Raven argues in Chapter 5 and elsewhere (1989), this classification distorts and obstructs efforts to assess significant educational achievements. Raven points out that not only are cognitive and affective outcomes treated as separate categories, but that what he calls the *conative* aspects of human behavior, those concerned with determination, persistence, and will, are inappropriately subsumed under "affective". A person, he points out, can enjoy doing something without being determined to see it through, and he or she can hate doing something, but still be determined to do it. He makes his argument focusing on the "ability to take initiative" which is generally acknowledged as a desirable educational outcome. He argues that taking initiative (which would be categorized as an "affective" outcome in the Bloom Taxonomy) is inseparable from intellectual or cognitive functioning, and from action:

> To take initiative successfully, people must be self-motivated. Self-starting people must be persistent and devote a great deal of time, thought, and effort to the activity. . . . The crucial point to be emphasized in attempting to clarify the nature of competence is that no one does any of these things unless he or she cares about the activity being undertaken. What a person values is therefore central. . . . What follows from this is that it is necessary to know an individual's values, interests, and preoccupations in order to assess his or her competencies. Important abilities demand time, energy, and effort. As a result, *people only display them when they are undertaking activities which are important to them* [Italics added] (Chapter 5, p. 89).

Raven goes on to argue that, if this analysis is correct, it does not make sense to attempt to assess separately cognitive, affective, and conative components of an activity. Affective and conative components are integral to the ability to cognize. "Not only do the three components interpenetrate if the

behavior in question—the taking of initiative—is to be successful, these components must be in balance. Determination exercised in the absence of understanding, and the converse, are unlikely to make for a competent performance."

The proposition that cognitive, affective, and conative aspects of human learning and development are inseparable is in sharp conflict with several accepted canons of traditional psychometry. It runs counter to the widespread practice of using one set of scales to assess values, attitudes, and beliefs and other independent scales to assess knowledge, skills, abilities, or competencies. Raven makes the intriguing suggestion that, if we are to assess such qualities as initiative, instead of trying to develop separate assessments which are difficult if not impossible to interpret, we need to develop indices which unify the cognitive, affective, and conative. He argues that development of all human capacities is highly contingent upon the opportunity structure (the social context), as well as on the learner's will, interest, and knowledge. In Chapter 5, Raven shows that it is technically possible to develop value-based indices that can do more justice both to the complexities of human qualities and capacities, and to how they are fostered and developed.

Assumption 4: The Need for Control from the Center. Testing and assessment procedures are forms of surveillance whose use is the superimposition of a power relationship. Criterion-referenced and standardized tests are sometimes criticized because they shape the school's curriculum and pedagogy. But the *raison d'etre* of *all* evaluative procedures in education, not only standardized and criterion-referenced tests, is to shape the educational process by exerting control over educational administrators, teachers, and/or students. Assessment procedures are inherently political, not only because whoever controls the assessment process shapes the curriculum pedagogy and ultimately the students' life chances, but also because particular forms of assessment promote *particular forms* of social control within the organization, while suppressing others.

My contention here is not that particular forms of organizational management and control inevitably follows from particular forms of assessment. Assessments are only one of many complex factors shaping how schools and school systems are governed. Rather, the claim is that the particular *form* of assessment is a key factor in producing particular forms of social control throughout the organization. In other words, the technology used in the assessment process, will encourage particular forms of management and human relationships within the organization while suppressing others.

The technology of mass administered standardized and criterion-referenced tests produces social relationships and management structures which are largely suited to exercising control from the center, that is, from the

central office by local or state educational authorities. Such tests provide virtually no information about what students are capable of doing or where they may need help. These tests produce relative rankings but little substantive information about what students know or can do which is useful to teachers, parents, prospective employers, or to students themselves for making programmatic or individual decisions. The psychometric technology only enables us to classify and rank order students (or teachers), and to constitute individuals as a "case," that is, as belonging to a class or category which possesses a particular set of objective characteristics (*e.g.* high, average, or low achievers, at risk students, etc.).

These tests are used primarily to facilitate what Michael Foucault (1979) calls *le regard*, (the gaze) or visibility to authority. Standardized tests and most criterion-referenced tests are particularly powerful forms of social control because they objectify the subject by reducing all human characteristics to a single number, thereby facilitating comparative rankings, and placing individuals into categories. These ranks and categories allow central office administrators to monitor and manage large numbers of students and teachers. Control exercised by such tests is not direct or overt, their effectiveness, rather, resides in the fact that those who are evaluated internalize or take into themselves the ranks and labels placed on them because these are presumably made by neutral, scientific instruments. Though individuals can and sometimes do resist these valuations of their capacities or achievements, the vast majority succumb because standardized and criterion-referenced test scores are the only educational currency accepted as scientific by the wider society.

Counter assumption: Assessment for Democratic Management Requires Dispersed Control. What should be emphasized for making a case for a contextual paradigm is that intrinsic to the use of standardized and criterion-referenced tests is a form of surveillance and exercise of power which is *unidirectional*. Central office administrators exert power over the everyday life and fate of students, teachers, and parents, who have no way of changing the system of assessment which controls them other than passive resistance or active subversion. While all forms of assessment, including any newer forms we might invent, represent a form of surveillance and constitute a means of control and an exercise of power, it is possible to alter the unidirectionality of control within the assessment system. That is to *re*-form the system of assessment in such a way that it disperses power, vesting it not only in administrative hands but also in the hands of teachers, students, parents and citizens of the community a particular school serves. If we are to have a system of public education supported by public funds, and governed by democratically elected bodies, then oversight by these bodies is essential. Some form of systematic assessment for holding educational institutions and the professionals who work in them accountable for their performance is necessary to monitor

expenditures, to insure that that professionals meet their responsibilities, do not exceed their authority, or violate the public trust or students' and parents' rights. But the exercise of power via the assessment process by central administrative authorities at the national, state, or district levels becomes coercive and oppressive without countervailing power over the assessment process exercised by teachers, parents, and students. From both experience and social scientific evidence, it is clear that good schools require a strong measure of autonomy by teachers, other school-level professionals, and participation by the local school-community. Without significant control over the assessment process at the school-level, teacher empowerment and school-based management is an illusion.

In Chapter 6 Elizabeth Adams and Tyrrell Burgess show that a system of assessment can be devised which vests significant power in the hands of central authorities, *and* in the hands of school-level professionals, parents, students, and the local school-community. Drawing upon their experience in the United Kingdom, they show how institutional arrangements and processes can be developed enabling the authorities at all levels to oversee the quality of schooling, to effect system-wide educational policies, and at the same time setting limits on the power of these authorities to trespass on the prerogatives of teachers, school heads, and students. In Chapter 8, I make an effort to extend their argument, and to show how such an effort could be adapted to fit the American experience.

Overview of the Book

A summary of the remaining chapters follows.

Chapter Two. Assessing Mathematics Competence and Achievement, by Thomas A. Romberg, defines and clarifies a conception of authentic achievement in mathematics and examines the validity of the commonly used instruments for assessing mathematics achievement. He concludes with a set of propositions to guide the development of new approaches to mathematics assessment and with an argument for the need to develop new approaches.

Chapter Three. The Assessment of Discourse in Social Studies, by Fred M. Newmann, suggests that a major aspect of social studies assessment should focus on the oral and written discourse that students produce on social topics. He addresses the questions of what discourse is and why it is an important indicator of student achievement in history and social studies. He concludes with his view of what experience and research suggest about the feasibility of this approach to assessment and with a discussion of how the assessment of discourse could provide meaningful and useful comparative indicators of student performance.

Chapter Four. In *The Nature of Authentic Academic Achievement* Fred M. Newmann and Doug A. Archbald argue for a particular view of "authentic" academic achievement, one that challenges the narrow conception of academic learning represented in virtually all current forms of standardized tests. They then examine the problems and implications of their view for assessing achievement.

Chapter Five. A Model of Competence, Motivation, and Behavior, and a Paradigm for Assessment, by John Raven, provides an argument for developing a new assessment paradigm, which is capable of assessing human capabilities and competencies. He proposes a model for measuring human capacities which unifies rather than separates the cognitive, affective, and conative aspects of human learning and development and, drawing upon his research, shows how such a model works in practice.

Chapter Six. In *Recognizing Achievement*, Elizabeth Adams and Tyrrell Burgess, drawing upon their extensive work, summarize an organizational model for a *system* of national assessment, one designed to provide information that enables teachers, students, parents, and public policy-makers to make wise and responsible choices and holds teachers and other professional educators accountable for their performance. Their proposed model, which has been implemented on a modest scale, underlines the interconnections between an assessment system and the way individual schools are structured and governed.

Chapter Seven. In *Approaches to Assessing Academic Achievement*, Doug A. Archbald and Fred M. Newmann include descriptions of efforts to develop alternatives to current forms of standardized and criterion-referenced tests. The efforts they report were generally initiated by school districts or individual schools and were not self-conscious efforts to apply a new theory of assessment. Rather, their development was driven by an effort to find practical ways of overcoming the limitations of conventional standardized and criterion-referenced tests. Though many are initial and partial efforts which would require considerable development before they could be considered for use on a wider scale, they do provide a rich set of possibilities for rethinking and reforming the assessment process.

Chapter Eight. In *Toward the Development of a New Science of Educational Testing and Assessment*, I challenge the basis of the claims that current forms of educational testing are objective or scientific. Drawing from earlier chapters and from several post-modernist and feminist writers, the chapter attempts to reconceptualize the assessment question and advance a case for a contextual paradigm. The chapter concludes with a discussion of the likely objections to the position taken and the prospects for change.

BEST COPY AVAILABLE

Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution

THOMAS M. HALADYNA SUSAN BOBBIT NOLEN NANCY S. HAAS

In the current climate of dissatisfaction with public education, the standardized achievement test score has been the operational definition for educational achievement, and raising test scores has been equated with educational improvement. The pressure to raise test scores has resulted in practices which pollute the inferences we make from these scores. We examine two major sources of test score pollution: (a) how public school personnel prepare students to take the standardized test and (b) nonstandard practices and conditions under which tests are administered. We also examine the apparent causes of this pollution and its effects on testing practices in American education.

Educational Researcher, Vol. 20, No. 5, pp. 2-7

The coin of the realm in public education in the United States is the standardized achievement test score. This score is universally and uncritically accepted by the public and many educators as a valid measure of educational accomplishment (Haertel & Calfee, 1983). In a review of testing practices related to standardized achievement testing, Haladyna, Haas, and Nolen (1989) listed 29 possible uses of standardized tests (see Table 1). These uses range from policy analysis at the national level to parental review of their child's achievement. For instance, test scores are used to rank states by the United States Department of Education in its annual "Report Card," and by legislators and other government officials to assess educational effectiveness of states and school districts. School boards and school district personnel use test scores to determine the effectiveness of their districts and schools within each district. Newspapers rank school districts by test scores to bemoan the failure of education. Test scores are used by some school district personnel to determine merit pay and to make other personnel decisions. Real estate agents use test scores to rate neighborhoods in terms of the "quality of schools."

Until recently, these test scores were used for a rather limited set of purposes. Scores were used to group students for instruction, evaluate and modify school district curricula, plan instruction, diagnose achievement deficits, place students into special programs (e.g., gifted, handicapped), and help parents understand the general achievement levels of their children. The considerable increase in the use of these test scores might be attributed to the onset of the "age of accountability" and an increased perceived need to evaluate education at virtually all units of analysis (i.e., individuals, classes, schools, school districts, states, and even the nation).

With the increased use of standardized achievement tests has come pressure to raise scores, which in turn leads to increased test score pollution. This pollution seriously affects the truthfulness of test score interpretations and calls into doubt the reaonableness of many of the uses listed in Table 1.

In this paper we will (a) define test score pollution, (b) describe the nature and extent of two major sources of test score pollution (student preparation and test administration practices), (c) discuss factors leading to polluting practices, and (d) suggest ways to combat test score pollution. But first, validity of test score use should be discussed as a context for the problem of test score pollution as it exists today.

The Validity Context

To validate a particular test use or interpretation, one should provide evidence supporting the truthfulness of that use. This fundamental principle is stated in standard 1.1 in the *Standards for Educational and Psychological Testing* (1985, p. 13)

Truthfulness

THOMAS M. HALADYNA *is Professor in the Division of Education and Human Services, Arizona State University, West Campus, PO Box 37100, Phoenix, AZ 85069-7100. His specializations include measurement, statistics, and research methods.* SUSAN BOBBIT NOLEN *is Assistant Professor in Educational Psychology, University of Washington, 322E Miller Call, DC-12, Seattle, WA 93195. Her specialization is learning and motivation.* NANCY S. HAAS *is Assistant Professor in the Education Unit, Arizona State University West, PO Box 37100, Phoenix AZ 85069-7100. Her area of specialization is instructional design and curriculum.*

Table 1
Consumers and Uses of Standardized Test Information

Consumer	Unit of analysis
National level	
Allocation of resources to programs and priorities	Nation, state
Federal program evaluation (e.g., Chapter I)	State, program
State legislature/state department of education	
Evaluate state's status and progress relevant to standards	State
State program evaluation	State, program
Allocation of resources	District, school
Public (lay persons, press, school board members, parents)	
Evaluate state's status and progress relevant to standards	District
Diagnose achievement deficits	Individual, school
Develop expectations for future success in school	Individual
School districts—central administrators	
Evaluate districts	District
Evaluate schools	Schools
Evaluate teachers	Classroom
Evaluate curriculum	District
Evaluate instructional programs	Program
Determine areas for revision of curriculum and instruction	District
School districts—building administrators	
Evaluate school	School
Evaluate teacher	Classroom
Grouping students for instruction	Individual
Placement into special programs	Individual
School districts—teachers	
Grouping students for instruction	Individual
Evaluating and planning the curriculum	Classroom
Evaluating and planning instruction	Classroom
Evaluating teaching	Classroom
Diagnosing achievement deficits	Classroom, individual
Promotion and graduation	Individual
Placement into special programs (e.g., gifted, handicapped)	Individual
Educational laboratories, centers, universities	
Policy analysis	All units
Evaluation studies	All units
Other applied research	All units
Basic research	All units

and more completely discussed in chapters on validity in the second and third editions of *Educational Measurement* (Cronbach, 1971; Messick, 1989). For the present purposes, the validity of any standardized achievement test score use is conceptualized in terms of a unified approach which embodies both content and construct considerations (Messick, 1989). In this validity context, any standardized achievement test score represents a generalized measure of accomplishment of school curricula (Waldrop et al., 1982). As Mehrens and Kaminski (1989) observed, the makers of standardized achievement tests hold that their tests are merely a sample from a broad achievement domain of knowledge and skills, a view consistent with the conception of validity we are using.

> Cronbach addressed this issue and its relationship to test preparation: Whenever it is critically important to master certain content, the knowledge that it will be tested produces a desirable concentration of effort. On the other hand, learning the answer to a set of questions is by no means the same as acquiring understanding of whatever topic that question represents (Cronbach, 1963, p. 681)

No single standardized achievement test represents a complete mapping of the content of the school achievement domain, nor is it so intended by its publishers. Indeed, many critics of standardized testing seek test use reform through the use of multiple indicators that better represent the complexity of school achievement. The December 1989 issue of the *Educational Researcher* is devoted to this idea of expanding the scope of educational achievement measurement.

Like the achievement domain, the causes of school achievement are varied and complex, a fact that further complicates the interpretation of achievement test scores. In Walberg's productivity model (1980) of school learning, for example, school achievement is a function of a variety of factors, only some of which are under the influence of schools. Schools can influence the quality and quantity of instruction,

BEST COPY AVAILABLE

motivation, and the learning environment, but they have little or no effect on family and home environment, maturity, and mental ability. Although researchers may disagree as to the relative degrees of influence of these factors or to the extent they interact, there is little disagreement that all are important.

Unfortunately, the consumers of school achievement test scores have often used test results without considering the complexity of achievement and its causes. Waldrop et al. (1982) maintain that standardized tests typically lack the "inference systems" necessary for many intended uses. For instance, erroneously attributing the level of achievement test scores to the influence of a single teacher, school, or school district grossly oversimplifies the nature of these scores. This misrepresentation seems to contribute to the growing fear and loathing that teachers and administrators feel toward standardized testing (Haas, Haladyna, & Nolen, 1989; Nolen, Haladyna, & Haas, 1989; Smith, with others, 1990).

There have been an increasing number of critics of this inappropriate use of test scores (Haertel, 1986; Madaus, 1988; Shepard, 1989). We believe that this increasing pressure to produce high test scores, combined with a lack of understanding of the complexities of achievement and its causes, has led to widespread practices that we have termed *test score pollution*.

Test Score Pollution

Test score pollution, a concept based on the work of Messick (1984), refers to factors affecting the truthfulness of a test score interpretaton. Specifically, pollution increases or decreases test performance *without connection to the construct* represented by the test, producing construct-irrelevant test score variance. Interpretations of these biased scores influence public opinion and policy and thus affect American education at all levels. There is reason to believe that the problem of test score pollution is pervasive in American education.

Three main sources of test score pollution are (a) the way schools and its personnel prepare students for tests, (b) test administration activities or conditions, and (c) exogenous factors representing forces beyond the control of schools and school personnel. Although this third factor is not part of the present discussion, we mention it briefly here. There are many factors believed to be causative agents of educational achievement. The reporting of test scores without acknowledging the influence of family, family mobility, economic environment, proficiency with the English language, and other such factors can lead both lay persons and educators to draw invalid inferences from these test scores. In this article however, we shall restrict the discussion of pollution to two factors under the control of school personnel.

Polluting Practices

Researchers have documented a number of activities aimed at preparing students for tests (Haladyna et al., 1989; Mehrens & Kaminski, 1989). These practices include (a) teaching test-taking skills, (b) promoting student motivation for the test, (c) developing a curriculum to match the test, (d) preparing teaching objectives to match the test, (e) presenting items similar to those presented on the test, (f) using commercial materials specifically designed to improve test performance, and (g) presenting before the test the actual items to be tested.

Polluting practices also occur during the actual administration of the tests. These include (a) "cleaning" answer sheets by darkening responses and erasing stray marks, (b) dismissing low-achieving students on test days, and (c) interfering with responses (e.g., giving hints or answers to students or altering response sheets). Other conditions exist in and around the testing situation that are known to influence test scores. These include students' anxiety, stress, fatigue, and motivation, as well as the pace of performance on the test as a function of the time available. This latter factor is also known to interact with anxiety and to yield predictably poor results (Matarazzo, 1972).

Table 2 provides our appraisal of the ethics of these practices. It must be emphasized that our views on which student preparation and test administration practices are ethical and unethical are not shared universally. Mehrens and Kaminski (1989) suggest that, to some extent, the acceptability of test preparation practices varies somewhat, depending on the purpose of testing. Our views, summarized in Table 2, are based on the assumption that scores are and will continue to be used to compare the educational effectiveness of teachers, administrators, classes, schools, districts, states, and nations.

Despite the fact that some practices may be considered ethical and others unethical, it must be noted that even ethical practices are polluting if they are unevenly ad-

Table 2
A Continuum of Test Preparation Activities

Test preparation activity	Degree of ethicality
Training in testwiseness skills	Ethical
Checking answer sheets to make sure that each has been properly completed	Ethical[1]
Increasing student motivation to perform on the test through appeals to parents, students, and teachers	Ethical
Developing a curriculum based on the content of the test	Unethical
Preparing objectives based on items on the test and teaching accordingly	Unethical
Presenting items similar to those on the test	Unethical
Using *Scoring High* or other score-boosting activities	Unethical
Dismissing low-achieving students on testing day to artificially boost test scores	Highly unethical
Presenting items verbatim from the test to be given	Highly unethical

[1]Ethical to the extent that the test publisher recommends it or to the extent that all schools, classes, and students being compared have the same service.

ministered within the unit of analysis being used for a particular test score interpretation. If one school district uses extensive test preparation programs for its students, is it valid to compare the test scores with those of a neighboring district where such test preparation does not take place?

Evidence of polluting test preparation and administration practices has been accumulating. Mehrens and Kaminski (1989) recently reviewed research on the variey of test preparation practices listed in Table 2. They concluded that indeed many of these tactics to improve test scores work quite well, and they also recognized that many of these activities tend to spuriously inflate test scores. Surveys and interviews conducted in Arizona, a state which mandates annual standardized achievement testing of all students in Grades 2 through 11, reveal widespread use of questionable and blatantly unethical practices to boost test scores (Haas et al., 1989; Nolen et al., 1989; Smith, with others, 1990).

Ethical Test Preparation Activities

Table 2 lists three classes of ethical activities to prepare students for standardized achievement tests. Training in testwiseness skills includes familiarizing students with the formats of answer sheets and test items and teaching general strategies for optimum performance on multiple-choice tests. Sarnacki (1979) provides a useful review of the methods and the effectiveness of these methods on test performance. Most (although not all) of our respondents indicated that such training was routine in their school or district (Haas et al., 1989; Nolen et al., 1989). Most respondents also reported various ethical strategies for increasing performance on these tests, such as demonstrating marking procedures (68.7%), sending notes home to parents about rest and nutrition (70.9%), encouraging attendance (92.6%), and discussing the purpose and importance of the tests (75.0%). It was also reported that some school districts check answer sheets very carefully for proper completion, whereas others do not. Variations in handling answer sheets mean that some schools or districts are disadvantaged by sloppy marking.

These surveys reveal that not all school districts practice the same kinds of ethical test preparation or in equal amounts. Although such practices should be encouraged and continued, as pointed out earlier in this article, the unequal application of these practices within a unit of analysis also pollutes test scores.

Unethical Test Preparation and Administration Activities

Both survey and interview responses indicated that certain unethical practices which inflate test scores without concurrently raising students achievement level were common in their school or district (Haas et al., 1989; Nolen et al., 1989). The elementary level teachers reported that 41.2% of them used commercial test preparation materials which, in addition to familiarizing students with test formats, teaches or reviews skills to be tested. Use of these materials may start several months prior to the test and may occur on a weekly or even daily basis. Mehrens and Kaminski (1989) have voiced the strongest objection to the ethics of using these test preparation programs, and their review is a most compelling argument against the continuation of this practice.

In addition to commercial packages, many educators reported that district curriculum and objectives had been "aligned" to the particular achievement test given (Haladyna et al., 1989). More informal alignment also occurs: Nolen et al., (1989) reported 8.5% of secondary level teachers and 10% of elementary school teachers teach students items from the current year's test.

We found evidence of nonstandard administration in both the survey and interview studies (Haas et al., 1989; Nolen et al., 1989). On the survey, 8% of the elementary school teachers reported deviations in reading test directions or increasing testing time, and more than a third stated that they weren't sure if they followed prescribed procedures exactly.

When does nonstandard administration become cheating? Some interviewees discussed helping students select answers for some items during the test, and 14.2% of survey respondents offered rewards for test completion. Some interviewees reported that low-achieving students were excused from taking the test by being dismissed or sent on field trips during test week. Such tactics will increase or ensure a much higher performance than deserved.

Undetected instances of cheating lead to false interpretations of test performance, one source of test score pollution. Collective damage occurs because of cheating. Teaching the items from the test or changing students' responses on the answer sheet leads to gross misinterpretations of student achievement. Less obviously unethical (and so perhaps more insidious) actions such as aligning curricula to the content of a particular achievement test, or extensive practice with alternate forms or commercial preparation packages, have a wider effect and can lead to misinterpretations of differences among schools and districts. By any reasonable standard, the extensiveness of score-polluting practices revealed in the reports reviewed here is staggering. As small as the percentages reported in various studies are, when multiplied by the millions of students taking these tests, the number of flawed results must be quite large.

A Climate for Test Score Pollution

Many of the current uses of test scores have been characterized as "high stakes" (Madaus, 1988) because tangible consequences depend upon test scores. When teachers' and principals' employment or salary advancement is linked to student performance on such tests, the stakes are indeed high.

As the number of uses for standardized achievement test scores has increased, so has the pressure to raise these scores. In a recent interview, George Madaus said:

> When the stakes are high, people are going to find ways to have test scores go up... The school will look better, but the skill levels will not necessarily be going up. You may have succeeded only in corrupting the inferences you wanted to make from the tests. (Brandt, 1989, p. 26)

This point has been made in a number of other contexts and times (e.g., Frederiksen, 1984).

Educators themselves may vary considerably on what preparation and administration practices they see as "cheating." Gonzales (1985, cited in Mehrens & Kaminski, 1989) surveyed teachers in one school district and found that 11% did not consider teaching students by using actual test items to be cheating. Thus some educators may be deluded into thinking that any practice that boosts scores is legitimate.

Cheating occurs in situations where the consequences of success or failure are important and where confidence in a successful outcome is not complete. It is more likely to occur when individuals are self-aware and comparing them-

436 BEST COPY AVAILABLE

selves to an external standard of performance (Malcolm & Ng, 1989). Earlier we established that the general public, the media, and politicians view standardized achievement test performance as an important indicator of academic effectiveness. It is clear that results have important consequences for individual teachers and administrators, as well as for districts, and that results (for schools or districts) are likely to be made public.

In high-stakes testing, many school personnel have an opportunity to "optimize" their students' performance, without necessarily increasing achievement. Although most educators who engage in what we have called unethical preparation and administration practices probably do not consider these activities as cheating, it is clear that the result is the same: polluted test scores.

Until there is serious reform in the way schools prepare students to take standardized achievement tests, test results will continue to misrepresent American public education and its accomplishments.

The fact that our respondents expressed considerable doubt as to the validity of scores obtained under the present system (a finding similar to data reported by Smith and others, 1990) suggests that many are aware of the problem of test score pollution and the external conditions that lead to polluting practices. As one anonymous teacher eloquently wrote (Haas et al., 1989):

> They [the tests] do not, in my opinion, reflect a student's progress. I resent the amount of teaching time it takes from my year to prep for the test. I feel that in order for their scores to be competitive, I must use that time, when it could be more effectively used for other learning purposes than test performance. (p. 96)

The majority of our participants stated that they felt pressure to increase student scores, but fewer than 20% felt the scores reflected a single year's learning. The extent of the pressure to raise scores varies from school to school and from district to district (a fact which in itself almost guarantees blatant attempts to boost test scores). It is clear from their comments that many teachers and administrators feel they are "under the gun" (Haas et al., 1989):

> It is a shame that people feel the need to evaluate one's achievement from marks on a bubble sheet. We as teachers are pressured to teach to the test. This is an absurd way to go about educating children. I feel that if I am pressured any more to do well on the TEST, then I will do everything I can to make sure my kids do well...even cheat....Is the real world a bubble sheet in which we base our decisions and our moral values? (p. 128)

The moral dilemma presented to teachers and administrators is more complex than it might seem. Smith, with others, (1990) wrote that teachers face the dissonance of matching their daily, intimate observation of each child with test results. This "interpretive context" is unique to the teacher of a classroom of children. Teachers report a consistent lack of belief in these test results and in particular with how the results are used (Haas et al., 1989; Nolen et al., 1989; Smith with others, 1990). Although teachers and administrators participate in efforts to raise test scores, as one elementary school principal commented, "we suffer a collective guilt in the process."

Back to the *Standards*

Messick (1975) stated that when one is considering the use of test scores for evaluation, not only the validity of the score interpretation but also the consequences of using the score in a particular way must be considered. The *Standards for Educational and Psychological Testing* makes clear that certain conditions must exist for public use of tests. The most vital of these conditions is that any test use must be supported by evidence attesting to the truthfulness of interpretations and the reasonableness of the use of these scores. The current overemphasis on standardized achievement test scores has created conditions under which they are badly polluted by test preparation activities, administration practices, and other conditions. Further, scores are reported without the context which may be critical to their interpretation. As a result, the uses of test results listed in Table 1 become questionable.

Until there is serious reform in the way schools prepare students to take standardized achievement tests, test results will continue to misrepresent American public education and its accomplishments. However, as long as test scores remain the single most important index of educational effectiveness, such reform is unlikely to take place. The educational research community must do its part to discourage polluting practices, as well as to educate the public about the problems inherent in overreliance upon standardized achievement test scores. More important, researchers, other educators, and policymakers must work together to develop means of evaluating educational effectiveness that accurately represent a school or district's progress toward a broad range of important educational goals. This work is beginning in a number of states, including Arizona, where our research was conducted.

As new evalution tools are developed, however, their use must also be examined carefully lest the problems we have discussed here continue. New forms of school achievement testing are still subject to the polluting influence we have described. We must look not only at what these new instruments tell us about student achievement but also at the ways in which students are prepared for and participate in the evaluation. As Messick, in his essay on meaning and values in measurement and evaluation (1975), writes,

> To judge the value of an outcome or end, one should understand the nature of the processes or means that led to that end, as Dewey (1939) emphasized in his principle of the means–end continuum: it's not just that the means are appraised in terms of the ends they lead to, but ends are appraised in terms of the means that produce them. (p. 963)

References

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus. *Educational Leadership, 46(7)*, 26–30.

Cronbach, L. J. (1963). Evaluation of course improvement. *Teachers College Record, 64*, 672–683.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507).

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.

Haas, N. S., Haladyna, T. M., & Nolen, S. B. (1989). *Standardized testing in Arizona: Interviews and written comments from teachers and administrators* (Tech. Rep. No. 89-3). Phoenix, AZ: Arizona State University West Campus.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational evaluation and policy analysis, 8*, 45–60.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*, 119–132.

Haladyna, T. M., Haas, N. S., Nolen, S. B. (1989). *Test score pollution* (Tech. Rep. No. 1). Phoenix, AZ: Arizona State University West Campus.

Madaus, G. F. (1988). The influence of testing on curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83–121). Chicago, IL: University of Chicago Press.

Malcolm, J., & Ng, S. H. (1989). Relationship of self-awareness to cheating on an external standard of competence. *Journal of Social Psychology, 129*, 391–395.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence*. New York, NY: Oxford University Press.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practices, 8*, 14–22.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215–237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13–104). Washington, DC: American Council on Education.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1989). *A survey of Arizona teachers and administrators on the uses and effects of state-mandated standardized achievement testing* (Tech. Rep. No. 89-2). Phoenix, AZ: Arizona State University West Campus.

Sarnacki, R. E. (1979). An examination of testwiseness in the cognitive test domain. *Review of Educational Research, 21*, 252–279.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46(7)*, 4–9.

Smith, M. L., with Edelsky, C., Draper, K., Rottenberg, C. & Cherland, M. (1990). *The role of testing in elementary schools*. Tempe, AZ: Arizona State University.

Walberg, H. J. (1980). A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.), *Perspectives on educational psychology*. Chicago & Berkeley, CA: National Society for the Study of Education & McCutchan Publishing.

Wardrop, J. L., Anderson, T. H., Hively, W., Hastings, C. N., Anderson, R. I., & Muller, K. E. (1982). A framework for analyzing the inference structure of educational achievement tests. *Journal of Educational Measurement, 19*, 1–18.

Note

This research was supported by the Arizona Department of Education and motivated by House Bill 2111, which was passed by the Arizona legislature. This law mandated careful study of the uses of standardized testing in Arizona schools.

The opinions expressed in this article are solely those of the authors and do not represent the official position of the Arizona Department of Education or the Arizona legislature.

Vygotsky's Psychology
A Biography of Ideas
Alex Kozulin

Alex Kozulin, translator of Vygotsky's work and distinguished Russian-American psychologist, has written the first major intellectual biography about Vygotsky's theories and their relationship to 20th-century Russian and Western intellectual culture. He traces Vygotsky's ideas to their origins in his early essays on literary criticism, Jewish culture, and the psychology of art, and he explicates brilliantly his psychological theory of language, thought, and development.
$29.95 cloth

Voices of the Mind
A Sociocultural Approach to Mediated Action
James V. Wertsch

James Wertsch, integrating a broad array of theoretical work, particularly that of Vygotsky and Bakhtin, outlines in this book an approach to mental functioning that stresses its inherent cultural, historical, and institutional context. A critical aspect of this approach is the cultural tools or "mediation means" that shape both social and individual processes. In considering how these mediation means—especially language—emerge in social history and in organizing the settings in which human beings are socialized, Wertsch achieves fresh insights into essential areas of human mental functioning that are typically unexplored or misunderstood.
$24.95 cloth

Harvard University Press
79 Garden Street
Cambridge, MA 02138 (617) 495-2480

45 BEST COPY AVAILABLE

Performance-Based Assessment and Educational Equity

LINDA DARLING-HAMMOND
Teachers College, Columbia University

The use of educational testing in the United States has been criticized for its inequitable effects on different populations of students. Many assume that new forms of assessment will lead to more equitable outcomes. Linda Darling-Hammond argues in this article, however, that alternative assessment methods, such as performance-based assessment, are not inherently equitable, and that educators must pay careful attention to the ways that the assessments are used. Some school reform strategies, for example, use assessment reform as a lever for external control of schools. These strategies, Darling-Hammond argues, are unlikely to be successful and the assessments are unlikely to be equitable because they stem from a distrust of teachers and fail to involve teachers in the reform processes.

Darling-Hammond argues instead for policies that ensure "top-down support for bottom-up reform," where assessment is used to give teachers practical information on student learning and to provide opportunities for school communities to engage in "a recursive process of self-reflection, self-critique, self-correction, and self-renewal." Ultimately, then, the equitable use of performance assessments depends not only on the design of the assessments themselves, but also on how well the assessment practices are interwoven with the goals of authentic school reform and effective teaching.

In recent years, the school reform movement has engendered widespread efforts to transform the ways in which students' work and learning are assessed in schools. These alternatives are frequently called performance-based or "authentic" assessments because they engage students in "real-world" tasks rather than multiple-choice tests, and evaluate them according to criteria that are important for actual performance in a field of work (Wiggins, 1989). Such assessments include oral presentations, debates, or exhibitions, along with collections of students' written products, videotapes of performances and other learning

This article was originally presented at the Ford Foundation Symposium, "Equity and Educational Testing and Assessment." Copyright © 1993 by the Ford Foundation. All rights reserved. Reprinted with permission. For permission to reproduce, contact the Ford Foundation.

Harvard Educational Review Vol. 64 No. 1 Spring 1994

BEST COPY AVAILABLE

occasions, constructions and models, and their solutions to problems, experiments, or results of scientific and other inquiries (Archbald & Newman, 1988). They also include teacher observations and inventories of individual students' work and behavior, as well as of cooperative group work (National Association for the Education of Young Children [NAEYC], 1988).

Much of the rationale for these initiatives is based on growing evidence that traditional norm-referenced, multiple-choice tests fail to measure complex cognitive and performance abilities. Furthermore, when used for decisionmaking, they encourage instruction that tends to emphasize decontextualized, rote-oriented tasks imposing low cognitive demands rather than meaningful learning. Thus, efforts to raise standards of learning and performance must rest in part on strategies to transform assessment practices.

In addition, efforts to ensure that *all* students learn in meaningful ways resulting in high levels of performance require that teachers know as much about students and their learning as they do about subject matter. However, teachers' understandings of students' strengths, needs, and approaches to learning are not well supported by external testing programs that send secret, secured tests into the school and whisk them out again for machine scoring that produces numerical quotients many months later. Authentic assessment strategies can provide teachers with much more useful classroom information as they engage teachers in evaluating how and what students know and can do in real-life performance situations. These kinds of assessment strategies create the possibility that teachers will not only develop curricula aimed at challenging performance skills, but that they will also be able to use the resulting rich information about student learning and performance to shape their teaching in ways that can prove more effective for individual students.

Recently, interest in alternative forms of student assessment has expanded from the classroom-based efforts of individual teachers to district and statewide initiatives to overhaul entire testing programs so that they become more performance-based. Major national testing programs, such as the National Assessment of Educational Progress and the College Board's Scholastic Assessment Tests (formerly the Scholastic Aptitude Tests), are also undergoing important changes. These programs are being redesigned so that they will increasingly engage students in performance tasks requiring written and oral responses in lieu of multiple-choice questions focused on discrete facts or decontextualized bits of knowledge.

However, proposals for assessment reform differ in several important ways: 1) in the extent to which they aim to broaden the roles of educators, students, parents, and other community members in assessment; 2) in the extent to which they aim to make assessment part of the teaching and learning process, and use it to serve developmental and educational purposes rather than sorting and screening purposes; 3) in the extent to which they anticipate a problem-based interdisciplinary curriculum or a coverage-oriented curriculum that maintains traditional subject area compartments for learning; and 4) in the extent to which they see assessment reform as part of a broader national agenda to improve and

equalize educational opportunities in schools. Some see assessment reform as part of a broader agenda to strengthen the national educational infrastructure (the availability of high-quality teachers, curriculum, and resources) and to equalize access so that all students start from an equal platform for learning. Others, however, view performance-based assessment as a single sledgehammer for change, without acknowledging other structural realities of schooling, such as vast inequalities in educational opportunities.

These differences in approaches to assessment reform predict very different consequences for the educational system, and dramatically different consequences for those who have been traditionally underserved in U.S. schools — students in poor communities, "minorities," immigrants, and students with distinctive learning needs. In this article, I argue in particular that ***changes in the forms of assessment are unlikely to enhance equity unless we change the ways in which assessments are used as well:*** from sorting mechanisms to diagnostic supports; from external monitors of performance to locally generated tools for inquiring deeply into teaching and learning; and from purveyors of sanctions for those already underserved to levers for equalizing resources and enhancing learning opportunities.

The extent to which educational testing serves to enhance teaching and learning and to support greater equality or to undermine educational opportunity depends on how a variety of issues are resolved. Among these are issues associated with the nature of assessment tools themselves:

- whether and how they avoid bias;
- how they resolve concerns about subjectivity versus objectivity in evaluating student work;
- how they influence curriculum and teaching.

A second set of issues has to do with whether and how assessment results are used to determine student placements and promotions, to reinforce differential curriculum tracking, or to allocate rewards and sanctions to teachers, programs, or schools.

A final set of issues concerns the policies and practices that surround the assessment system and determine the educational opportunities available to students to support their learning. A fundamental question is whether assessment systems will support better teaching and transform schooling for traditionally underserved students or whether they will merely reify existing inequities. This depends on the extent to which they promote equity in the allocation of resources for providing education, supports for effective teaching practices, and supports for more widespread school restructuring.

Motivations for Assessment Reform

The current movement to change U.S. traditions of student assessment in large-scale and systemic ways has several motivations. One is based on the recognition that assessment, especially when it is used for decisionmaking purposes, exerts

powerful influences on curriculum and instruction. It can "drive" instruction in ways that mimic not only the content, but also the format and cognitive demands of tests (Darling-Hammond & Wise, 1985; Madaus, West, Harmon, Lomax, & Viator, 1992). If assessment exerts these influences, many argue, it should be carefully shaped to send signals that are consistent with the kinds of learning desired and the approaches to curriculum and instruction that will support such learning (Cohen & Spillane, 1992; O'Day & Smith, 1993).

A second and somewhat related motive for systemic approaches to assessment reform stems from the belief that if assessment can exert powerful influences on behavior, it can be used to change school organizational behavior as well as classroom work. The idea of using assessment as a lever for school change is not a new one: many accountability tools in the 1970s and 1980s tried to link policy decisions to test scores (Linn, 1987; Madaus, 1985; Wise, 1979). Unfortunately, these efforts frequently had unhappy results for teaching and learning generally, and for schools' treatment of low-scoring students in particular. Research on these initiatives has found that test-based decisionmaking has driven instruction toward lower order cognitive skills. This shift has created incentives for pushing low scorers into special education, consigning them to educationally unproductive remedial classes, holding them back in the grades, and encouraging them to drop out (Allington & McGill-Franzen, 1992; Darling-Hammond, 1991, 1993; Koretz, 1988; Shepard & Smith, 1988; Smith, 1986). In addition, school incentives tied to test scores have undermined efforts to create and sustain more inclusive and integrated student populations, as schools are punished for accepting and keeping students with special needs and are rewarded for keeping such students out of their programs through selective admissions and transfer policies. Those with clout and means "improve" education by manipulating the population of students they serve (Smith, 1986). Schools serving disadvantaged students find it increasingly hard to recruit and retain experienced and highly qualified staff when the threat of punishments for low scores hangs over them. Thus, such policies exacerbate rather than ameliorate the unequal distribution of educational opportunity.

Nonetheless, a variety of proposals have recently been put forth that involve the use of mandated performance-based assessments as external levers for school change (Commission on Chapter I, 1992; Hornbeck, 1992; O'Day & Smith, 1993). Even those who do not endorse such proposals share the view that assessment can promote change. Other proposals, raised from a different philosophical vantage point and envisioning different uses of assessment, suggest the use of alternative classroom-embedded assessments as internal supports for school-based inquiry (Darling-Hammond & Ascher, 1990; Wolf & Baron, in press).

A third reason for assessment reform addresses concerns about equity and access to educational opportunity. Over many decades, assessment results have frequently been used to define not only teaching, but also students' opportunities to learn. As a tool for tracking students into different courses, levels, and kinds of instructional programs, testing has been a primary means for limiting or expanding students' life choices and their avenues for demonstrating competence. Increasingly, these uses of tests are recognized as having the unin-

tended consequence of limiting students' access to further learning opportunities (Darling-Hammond, 1991; Glaser, 1990; Oakes, 1985).

Some current proposals for performance-based assessment view these new kinds of tests as serving the same screening and tracking purposes as more traditional tests (Commission on the Skills of the American Workforce, 1990; Educate America, 1991; National Center on Education and the Economy, 1989). The presumption is that more "authentic" assessments will both motivate and sort students more effectively. Others see a primary goal of assessment reform as transforming the purposes and uses of testing as well as its form and content. They argue for shifting from the use of assessment as a sorting device to its use as a tool for identifying student strengths and needs so that teachers can adapt instruction more successfully (Darling-Hammond, Ancess, & Falk, in press; Glaser, 1981, 1990; Kornhaber & Gardner, 1993). Given the knowledge now available for addressing diverse learning needs and the needs of today's society for a broadly educated populace, the goals of education — and assessment — are being transformed from deciding who will be permitted to become well-educated to helping ensure that everyone will learn successfully.

Clearly, the current press to reform assessment entails many motivations and many possible consequences, depending on decisions that are made about 1) the nature of the "new" assessments; 2) the ways in which they are used; and 3) the companion efforts (if any) that accompany them to actually improve education in the schools.

In this article I outline the range of equity issues that arise with respect to testing generally, and with respect to proposals for the development of new "authentic" assessments specifically. I argue that the outcomes of the current wave of assessment reforms will depend in large measure on the extent to which assessment developers and users:

- focus on both the quality and fairness of assessment strategies;
- use assessments in ways that serve teaching and learning, rather than sorting and selecting;
- develop policies that are congruent with (and respectful of) these assessment goals, as well as with assessment strategies and limitations;
- embed assessment reform in broader reforms to improve and equalize access to educational resources and opportunities;
- support the professional development of teachers along with the organizational development of schools, so that assessment is embedded in teaching and learning, and is used to inform more skillful and adaptive teaching that enables more successful learning for all students.

Uses and Consequences of Testing

Historical Perspectives

For over one hundred years, standardized testing has been a tool used to exert control over the schooling process and to make decisions about educational

entitlements for students. Testing proved a convenient instrument of social control for those superintendents in the late nineteenth century who sought to use tests as a means for creating the "one best system" of education (Tyack, 1974). It also proved enormously useful as a means of determining how to slot students for more and less rigorous (and costly) curricula when public funding of education and compulsory attendance vastly increased access to schools in the early twentieth century.

Given the massive increase in students, the limits of public budgets, and the relatively meager training of teachers, strategies were sought to codify curriculum and to group students for differential instruction. IQ tests were widely used as a measure of educational input (with intelligence viewed as the "raw material" for schooling) to sort pupils so they could be efficiently educated according to their future roles in society (Cremin, 1961; Cubberly, 1919; Watson, in press). Frequently, they were used to exclude students from schooling opportunities altogether (Glaser, 1981).

Though many proponents argued that the use of these tests as a tool for tracking students would enhance social justice, the rationales for tracking — like those for using scores to set immigration quotas into the United States — were often frankly motivated by racial and ethnic politics. Just as Goddard's 1912 data — "proving" that 83 percent of Jews, 80 percent of Hungarians, 79 percent of Italians, and 87 percent of Russians were "feebleminded" — were used to justify low immigration quota for those groups (Kamin, 1974), so did Terman's test data "prove" that "[Indians, Mexicans, and Negroes] should be segregated in special classes. . . . They cannot master abstractions, but they can often be made efficient workers" (Terman, cited in Oakes, 1985, p. 36). Presumptions like these reinforced racial segregation and differential learning opportunities.

Terman found many inequalities in performance among groups on his IQ test, which was adapted from Binet's work in France. Most, but not all of them, seemed to confirm what he, and presumably every "intelligent" person, already knew: that various groups were inherently unequal in their mental capacities. However, when girls scored higher than boys on his 1916 version of the Stanford-Binet, he revised the test to correct for this apparent flaw by selecting items to create parity among genders in the scores (Mercer, 1989). Other inequalities — between urban and rural students, students of higher and lower socioeconomic status, native English speakers and immigrants, Whites and Blacks — did not occasion such revisions, since their validity seemed patently obvious to the test-makers.

The role of testing in reinforcing and extending social inequalities in educational opportunities has by now been extensively researched (Gould, 1981; Kamin, 1974; Mercer, 1989; Oakes, 1985; Watson, in press) and widely acknowledged. It began with the two fallacies Gould describes: the fallacy of reification, which allowed testers to develop and sell the abstract concept of intelligence as an innate, unitary, measurable commodity; and the fallacy of ranking, which supported the development of strategies for quantifying intelligence in ways that would allow people to be arrayed in a single series against each other (Gould,

1981). These two fallacies — recently debunked (though not yet dismantled) by understandings that intelligence has many dimensions (Gardner, 1983; Sternberg, 1985) — were made more dangerous by the social uses of testing as a tool for allocating educational and employment benefits rather than as a means for informing teaching and developing talents.

Negative Consequences of Standardized Testing

Current standardized tests are widely criticized for placing test-takers in a passive, reactive role (Wigdor & Garner, 1982), rather than one that engages their capacities to structure tasks, produce ideas, and solve problems. Based on outmoded views of learning, intelligence, and performance, they fail to measure students' higher order cognitive abilities or to support their capacities to perform real-world tasks (Resnick, 1987a; Sternberg, 1985).

In a seminal paper on the past, present, and future of testing, Glaser (1990) makes an important distinction between testing and assessment. These two kinds of measurement have different purposes and different social and technical histories. Glaser describes testing as aimed at selection and placement: it attempts to predict success at learning by "measur[ing] human ability prior to a course of instruction so that individuals can be appropriately placed, diagnosed, included or excluded" (p. 2). Assessment, on the other hand, is aimed at gauging educational outcomes: it measures the results of a course of learning. What is important for testing is the instrument's predictive power rather than its content. What is important for assessment is the content validity of an approach — its ability to describe the nature of performance that results from learning.

Recently, another validity construct has emerged: ***consequential validity,*** which describes the extent to which an assessment tool ***and the ways in which it is used*** produce positive consequences both for the teaching and learning process and for students who may experience different educational opportunities as a result of test-based placements (Glaser, 1990; Shepard, 1993). This emerging validity standard places a much heavier burden on assessment developers and users to demonstrate that what they are doing works to the benefit of those who are assessed and to the society at large. The emergence of this standard has led many educators and researchers to question test-based program placements for students and to press for forms of assessment that can support more challenging and authentic forms of teaching and learning. Some test developers are just beginning to understand that the criteria against which their products are being evaluated are changing.

For most of this century, much of the energy of U.S. measurement experts has been invested in developing tests aimed at ranking students for sorting and selecting them into and out of particular placements. Standardized test developers have devoted much less energy to worrying about the properties of these instruments as reflections of — or influences on — instruction (Wigdor & Garner, 1982). As a consequence, the tests generally do not reflect the actual tasks educators and citizens expect students to be able to perform, nor do they stimulate forms of instruction that are closely connected to development of perform-

ance abilities. Similarly, to date, though awareness levels are heightened, virtually no attention has yet been paid to the consequences of test-based decisions in policy discussions about developing new assessment systems.

These shortcomings of U.S. tests were less problematic when they were used as only one source of information among many other kinds of information about student learning, and when they were not directly tied to decisions about students and programs. However, as test scores have been used to make important educational decisions, their flaws have become more damaging. As schools have begun to "teach to the tests," the scores have become ever poorer assessments of students' overall abilities, because class work oriented toward recognizing the answers to multiple-choice questions does not heighten students' proficiency in aspects of the subjects that are not tested, such as analysis, complex problem-solving, and written and oral expression (Darling-Hammond & Wise, 1985; Haney & Madaus, 1986; Koretz, 1988).

As the National Assessment of Educational Progress (NAEP) found: "Only 5 to 10 percent of students can move beyond initial readings of a text; most seem genuinely puzzled at requests to explain or defend their points of view." The NAEP assessors explained that current methods of testing reading require short responses and lower level cognitive thinking, resulting in "an emphasis on shallow and superficial opinions at the expense of reasoned and disciplined thought, . . . [thus] it is not surprising that students fail to develop more comprehensive thinking and analytic skills" (NAEP, 1981, p. 5).

During the 1970s, when test-oriented accountability measures were instituted in U.S. schools, there was a decline in public schools' use of teaching methods appropriate to the teaching of higher order skills, such as research projects and laboratory work, student-centered discussions, and the writing of essays or themes (National Center for Education Statistics [NCES], 1982, p. 83). Major studies by Boyer (1983), Goodlad (1984), and Sizer (1985) documented the negative effects of standardized testing on teaching and learning in high schools, while the disadvantage created for U.S. students by the rote learning stressed in U.S. standardized tests has been documented in international studies of achievement (McKnight et al., 1987).

The effects of basic skills test misuse have been most unfortunate for the students they were most intended to help. Many studies have found that students placed in the lowest tracks or in remedial programs — disproportionately low-income and minority students — are most apt to experience instruction geared only to multiple-choice tests, working at a low cognitive level on test-oriented tasks that are profoundly disconnected from the skills they need to learn. Rarely are they given the opportunity to talk about what they know, to read real books, to write, or to construct and solve problems in mathematics, science, or other subjects (Cooper & Sherk, 1989; Davis, 1986; Oakes, 1985; Trimble & Sinclair, 1986). In short, they have been denied the opportunity to develop the capacities they will need for the future, in large part because commonly used tests are so firmly pointed at educational goals of the past.

Thus, the quality of education made available to many students has been undermined by the nature of the testing programs used to monitor and shape

their learning. If new performance-based assessments point at more challenging learning goals for all students, they may ameliorate some of this source of inequality. However, this will be true only to the extent that teachers who serve these students are able to teach in the ways demanded by the assessments — that is, in ways that support the development of higher order thinking and performance skills and in ways that diagnose and build upon individual learners' strengths and needs.

The Uses of Assessment Tools in Decisionmaking

As noted earlier, testing policies affect students' opportunities to learn in other important ways. In addition to determining whether students graduate, tests are increasingly used to track students and to determine whether they can be promoted from one grade to the next. Research suggests that both practices have had harmful consequences for individual students and for U.S. achievement generally. If performance-based assessments are used for the same purposes as traditional tests have been, the outcomes for underserved students are likely to be unchanged.

Tracking In the United States, the process of tracking begins in elementary schools with the designation of instructional groups and programs based on test scores, and becomes highly formalized by junior high school. The result of this practice is that challenging curricula are rationed to a very small proportion of students. Consequently, few U.S. students ever encounter the kinds of curricula that most students in other countries typically experience (McKnight et al., 1987). As Oakes (1986) notes, these assignments are predictable:

> One finding about placements is undisputed. . . . Disproportionate percentages of poor and minority youngsters (principally black and Hispanic) are placed in tracks for low-ability or non-college-bound students (NCES, 1985; Rosenbaum, 1980); further, minority students are consistently underrepresented in programs for the gifted and talented. (College Board, 1985, p. 129)

Students placed in lower tracks are exposed to a limited, rote-oriented curriculum and ultimately achieve less than students of similar aptitude who are placed in academic programs or untracked classes. Furthermore, these curricular differences explain much of the disparity between the achievement of White and minority students and between those of higher and lower income levels (Lee & Bryk, 1988; Oakes, 1985). In this way, the uses of tests have impeded rather than supported the pursuit of high and rigorous educational goals for all students.

Grade Retention In addition, some U.S. states and local districts have enacted policies requiring that test scores be used as the sole criterion for decisions about student promotion from one grade to the next. Since the student promotion policies were enacted, a substantial body of research has demonstrated that the effects of this kind of test-based decisionmaking are much more negative than positive. When students who were retained in grade are compared to students of equal achievement levels who were promoted, the retained students are con-

sistently behind on both achievement and social-emotional measures (Holmes & Matthews, 1984; Shephard & Smith, 1986). As Shephard and Smith put it, "Contrary to popular beliefs, repeating a grade does *not* help students gain ground academically and has a negative impact on social adjustment and self-esteem" (1986, p. 86).

Furthermore, the practice of retaining students is a major contributor to increased dropout rates. Research suggests that being retained increases the odds of dropping out by 40 to 50 percent. A second retention nearly doubles the risk (Mann, 1987; see also Carnegie Council on Adolescent Development, 1989; Massachusetts Advocacy Center, 1988; Wehlage, Rutter, Smith, Lesko, & Fernández, 1990). Thus, the policy of automatically retaining students based on their test-score performance has actually produced lower achievement for these students, lower self-esteem, and higher dropout rates for them and for the nation.

Graduation Perhaps the ultimate test-related sanction for students is denying a diploma based on a test score. The rationale for this practice is that students should show they have mastered the "minimum skills" needed for employment or future education in order to graduate. The assumption is that tests can adequately capture whatever those skills are. While this appears plausible in theory, it is unlikely in reality, given the disjunction between multiple-choice tests of decontextualized bits of information and the demands of real jobs and adult tasks (Bailey, 1989; Carnevale, Gainer, & Meltzer, 1989; Resnick, 1987b). In fact, research indicates that neither employability nor earnings are significantly affected by students' scores on basic skills tests, while chances of employment and welfare dependency are tightly linked to graduation from high school (Eckland, 1980; Gordon & Sum, 1988; Jaeger, 1991). Thus, the use of tests as a sole determinant of graduation imposes heavy personal and societal costs, without obvious social benefits.

Rewards and Sanctions Finally, a few states and districts have also tried to use student test scores to allocate rewards or sanctions to schools or teachers. President Bush's proposal for a National Test included a suggestion to allocate some federal funds based on schools' scores on the "American Achievement Tests" (U.S. Department of Education, 1991). An independent commission on Chapter I has recently proposed, over the formal dissent of a number of its members, a rewards and sanctions system for Chapter I programs based on aggregate "performance-based" test scores (Commission on Chapter I, 1992). An analogous policy proposal has been enacted, though not yet implemented, for use with performance-based tests in the state of Kentucky. There, all schools that do not show specified percentage increases in student achievement scores each year will automatically suffer sanctions, which may include actions against staff. Those that meet the standards will be financially rewarded (Legislative Research Commission, 1990, p. 21).

Oblivious to the fact that schools' scores on any measure are sensitive to changes in the population of students taking the test, and that such changes can be induced by manipulating admission, dropouts, and pupil classifications, the

policy will create and sustain a wide variety of perverse incentives, regardless of whether the tests are multiple-choice or performance-oriented. Because schools' aggregate scores on any measure are sensitive to the population of students taking the test, the policy creates incentives for schools to keep out students whom they fear may lower their scores — children who are handicapped, limited English speaking, or from educationally disadvantaged environments. Schools where average test scores are used for making decisions about rewards and sanctions have found a number of ways to manipulate their test-taking population in order to inflate artificially the school's average test scores. These strategies include labelling large numbers of low-scoring students for special education placements so that their scores won't "count" in school reports, retaining students in grade so that their relative standing will look better on "grade-equivalent" scores, excluding low-scoring students from admission to "open enrollment" schools, and encouraging such students to leave schools or drop out (Allington & McGill-Franzen, 1992; Darling-Hammond, 1991, 1993; Koretz, 1988; Shepard & Smith, 1988; Smith, 1986).

Smith explains the widespread engineering of student populations that he found in his study of a large urban school district that used performance standards as a basis for school level sanctions:

> Student selection provides the greatest leverage in the short-term accountability game. . . . The easiest way to improve one's chances of winning is (1) to add some highly likely students and (2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric. (1986, pp. 30–31)

Needless to say, this kind of policy that rewards or punishes schools for aggregate test scores creates a distorted view of accountability, in which beating the numbers by playing shell games with student placements overwhelms efforts to serve students' educational needs well. Equally important, these policies further exacerbate existing incentives for talented staff to opt for school placements where students are easy to teach, and school stability is high. Capable staff are less likely to risk losing rewards or incurring sanctions by volunteering to teach where many students have special needs and performance standards will be more difficult to attain. This compromises even further the educational chances of disadvantaged students, who are already served by a disproportionate share of those teachers who are inexperienced, unprepared, and under qualified.

Applying sanctions to schools with lower test score performance penalizes already disadvantaged students twice over: having given them inadequate schools to begin with, society will now punish them again for failing to perform as well as other students attending schools with greater resources and more capable teachers. This kind of reward system confuses the quality of education offered by schools with the needs of the students they enroll; it works against equity and integration, and against any possibilities for fair and open school choice, by discouraging good schools from opening their doors to educationally needy students. Such a reward structure places more emphasis on score manipulations

and student assignments or exclusions than on school improvement and the development of more effective teaching practices.

Policies for Building an Equitable System

Improving Teacher Capacity

Because this nation has not invested heavily in teacher education and professional development, the capacity for a more complex, student-centered approach to teaching is not prevalent throughout the current teaching force. Furthermore, because teacher salaries and working conditions are inadequate to ensure a steady supply of qualified teachers in poor districts, low-income and minority students are routinely taught by the least experienced and least prepared teachers (Darling-Hammond, 1991; Oakes, 1990). Differences in achievement between White and minority students can be substantially explained by unequal access to high-quality curriculum and instruction (Barr & Dreeben, 1983; College Board, 1985; Darling-Hammond & Snyder, 1992a; Dreeben, 1987; Dreeben & Barr, 1987; Dreeben & Gamoran, 1986; Oakes, 1990).

From a policy perspective, perhaps the single greatest source of educational inequity is this disparity in the availability and distribution of highly qualified teachers (Darling-Hammond, 1990). Providing equity in the distribution of teacher quality will be required before changes in assessment strategies result in more challenging and effective instruction for currently underserved students. This, in turn, requires changing policies and long-standing incentive structures in education so that shortages of well-prepared teachers are overcome, and schools serving poor and minority students are not disadvantaged by lower salaries and poorer working conditions in the bidding war for good teachers. Fundamental changes in school funding are essential to this task. Since revenues in poor districts are often half as great as those in wealthy districts, state aid changes that equalize district resources are the first step toward ensuring access to qualified teachers (Darling-Hammond, in press).

This crucial equity concern is finally gaining some attention in the rush to improve schools by testing. The recent report of the National Council on Education Standards and Testing (NCEST), while arguing for national performance standards for students, acknowledged the importance of "school delivery standards" for educational improvements to occur. The Council's Standards Task Force noted:

> If not accompanied by measures to ensure equal opportunity to learn, national content and performance standards could help widen the achievement gap between the advantaged and the disadvantaged in our society. If national content and performance standards and assessment are not accompanied by clear school delivery standards and policy measures designed to afford all students an equal opportunity to learn, the concerns about diminished equity could easily be realized. Standards and assessments must be accompanied by policies that provide access for all students to high quality resources, including appropriate instructional materials and well-prepared teachers. High content and performance standards can be used to chal-

lenge all students with the same expectations, but high expectations will only result in common high performance if all schools provide high quality instruction designed to meet the expectations. (NCEST, 1992, pp. E12–E13)

Delivery standards make clear that the governmental agencies that are imposing standards upon students are simultaneously accepting responsibility for ensuring that students will encounter the opportunities necessary for their success (Darling-Hammond, 1993). Though this may seem a straightforward prerequisite for making judgments about students or schools, it marks an entirely different approach to accountability in U.S. education than the one that has predominated for most of the last two decades and is widespread today. Earlier approaches to outcomes-based accountability legislated minimum competency tests and sometimes punished schools or students with low scores without attempting to correct the resource disparities that contributed to poor performance in the first place.

Ensuring that all students have adequate opportunities to learn requires enhancing the capacity of all teachers — their knowledge of students and subjects, and their ability to use that knowledge — by professionalizing teaching. This means that teacher education policies must ensure that *all* teachers have a stronger understanding of how children learn and develop, how assessment can be used to evaluate what they know and how they learn, how a variety of curricular and instructional strategies can address their needs, and how changes in school and classroom organization can support their growth and achievement.

Such teacher capacities are also important for supporting the promise of authentic assessment to enable richer, more instructionally useful forms of evaluation that are also fair and informative. A major reason for the advent of externally controlled highly standardized testing systems has been the belief that teachers could not be trusted to make sound decisions about what students know and are able to do. The presumed "objectivity" of current tests derives both from the lack of reliance upon individual teacher judgment in scoring and from the fact that test-takers are anonymous to test-scorers (hence, extraneous views about the student do not bias scoring).

Of course, many forms of bias remain, as the choice of items, responses deemed appropriate, and content deemed important are the product of culturally and contextually determined judgments, as well as the privileging of certain ways of knowing and modes of performance over others (García & Pearson, in press; Gardner, 1983; Sternberg, 1985; Wigdor & Garner, 1982). And these forms of bias are equally likely to plague performance-based assessments, as the selection of tasks will rest on cultural and other referents, such as experiences, terms, and exposures to types of music, art, literature, and social experiences that are differentially accessible to test-takers of different backgrounds.

If assessment is to be used to open up as many opportunities as possible to as many students as possible, it must address a wide range of talents, a variety of life experiences, and multiple ways of knowing. Diverse and wide-ranging tasks that use many different performance modes and that involve students in choosing ways to demonstrate their competence become important for this goal (Gor-

don, no date; Kornhaber & Gardner, 1993). Substantial teacher and student involvement in and control over assessment strategies and uses are critical if assessment is to support the most challenging education possible for every student, taking full account of his or her special talents and ways of knowing. As Gordon puts it:

> The task is to find assessment probes which measure the same criterion from contexts and perspectives which reflect the life space and values of the learner. . . . Thus options and choices become a critical feature in any assessment system created to be responsive to equity, just as processual description and diagnosis become central purposes. (no date, pp. 8–9)

The objective of maintaining high standards with less standardization will demand teachers who are able to evaluate and eliminate sources of unfair bias in their development and scoring of instructionally embedded assessments, and who can balance subjectivity and objectivity, using their subjective knowledge of students appropriately in selecting tasks and assessment options while adhering to common, collective standards of evaluation. These same abilities will be crucial for other assessment developers. In many respects, even greater sensitivity to the sources of bias that can pervade assessment will be needed with forms that frequently eliminate the anonymity of test-takers, drawing more heavily on interpersonal interaction in tasks and on observations on the part of teachers.

up to here

"Top-Down Support for Bottom-Up Reform"

The need for greater teacher knowledge and sensitivity in developing and using authentic assessments in schools will cause some to argue that they should not be attempted; that externally developed and scored "objective" tests are safer for making decisions because local judgement is avoided. However, the argument for authentic assessments rests as much on a changed conception of the *uses* of assessment as on the *form* in which assessment occurs. Rather than being used largely to determine how students rank against one another on a single, limited dimension of performance so as to determine curriculum or school placements of various kinds, many reformers hope that assessment can be used to *inform and improve* teaching and learning.

In this view, assessment should be integrally connected to the teaching and learning process so that students' strengths and needs are identified, built upon, and addressed. School-wide assessments should continually inform teachers' collective review of their practice so that improvements in curriculum, instruction, and school organization are ongoing. Thus, students should actually *learn* more as a result of assessment, rather than being more precisely classified, and schools should be able to inquire into and improve their practices more intelligently, rather than being more rigidly ranked. Assessment should increase the overall amount of learning and good practice across all schools, rather than merely measuring how much of a nonexpanding pool of knowledge is claimed by different students and schools.

If authentic assessment is to realize its potential as a tool for school change, however, policies must enable assessments to be used as a vehicle for student,

Wiggins

Curriculum Development, National Council of Teachers of English, National Council of Teachers of Mathematics, American Federation of Teachers, and American Association of School Administrators, and other national educational organizations.

Wiggins's current work as a researcher and consultant on performance-based curricula and assessment is grounded in many years of teaching and coaching. His career spanned fourteen years and three different disciplines (English, social studies, and mathematics) at the secondary level. He is also known for his pioneering work in the teaching of philosophy at the secondary level. Wiggins has coached four interscholastic sports for boys and girls: soccer, cross-country, track, and baseball. He is married to Holly Houston; they have two young sons, Justin and Ian.

1 Introduction: Assessment and the Morality of Testing

> People write the history of experiments on those born blind, on wolf-children, or those under hypnosis. But who will write the more general, more fluid, but also more determinant history of the "examination"—its rituals, its methods, its characters and their roles, its play of questions and answers, its systems of marking and classification? For in this slender technique are to be found a whole domain of knowledge, a whole type of power.
>
> —Michel Foucault[1]

Consider, for a moment, one unorthodox "examination" strategy that now exists within our educational world. In this form of assessment, the challenges are not at all standardized; indeed, they are by design fully personalized, allowing each student free rein as to topic, format, and style. And although students are not subject to uniform tasks required of all, no one thinks this lack of uniformity is odd or "soft." Contrary to common practice, this test is never secret: the assessment is centered on students' intellectual interests and the thoughtfulness with which those ideas are pursued. Students are assessed on how well knowledge and interest are crafted into products and performances of their own design. No student must "earn" this right to create projects and works of one's choosing; this is an assumed right.

The setting for such assessment is mainstream, not "alternative." Nonetheless, the schedule in support of such assessment is out

of the ordinary, designed to suit the learner's, not the teacher's pace; each student is assessed only when ready. Instead of having an adversarial relationship, teacher and student are allies. The teacher is the student's guide through the challenges of the assessment, not an enemy to be "psyched out." The assessor is in fact *obligated* to understand the student's point of view in order to validly assess the student's grasp of things—a far cry from the typical "gotcha!" test.

Perhaps by now you have guessed the locale of such scenes: many kindergartens and graduate schools throughout our country. At both extremes of the school career, we deemphasize one-shot uniform testing in favor of a careful assessment, from different perspectives, of the student's own projects. We focus more on the student's ability to extend or play with ideas than on the correctness of answers to generic questions. Each piece of work, be it a drawing or a dissertation, is examined—often through dialogue—for what it reveals about the learner's habits of mind and ability to create meaning, not his or her "knowledge" of "facts." At the beginning and end of formal education, we understand that intellectual accomplishment is best judged through a "subjective" but rigorous interaction of mind and mind. Since performance based on one's own inquiry is being measured, test "security" is a moot issue. The essential aim behind this kind of tactful examining can be expressed as a question: Do the student's particular ideas, arguments, and products *work*—that is, do the work-products effectively and gracefully achieve the student's intention?

During the bulk of schooling, unfortunately, the story is far different. From first grade through at least the beginning (and often the end) of the undergraduate years in college, standardized and short-answer tests—and the mentality that they promote—are dominant. Students are tested not on the way they use, extend, or criticize "knowledge" but on their ability to generate a superficially correct response on cue. They are allowed one attempt at a test that they know nothing about until they begin taking it. For their efforts, they receive—and are judged by—a single numerical score that tells them little about their current level of progress and gives them no help in improving. The result is, as Lauren and Daniel Resnick, researchers and heads of the New Standards Project, have written,

that American students are the "most tested but least examined" in the world.[2]

Better Assessment, Not Better Testing

But "read me well!" as Nietzsche asked of his readers in the preface to *Dawn*. This book is not a (predictable) critique of multiple-choice tests. On the contrary, *we have the tests we deserve*, including the tests that teachers design and give. The stubborn problems in assessment reform have to do with a pervasive thoughtlessness about testing and a failure to understand the relationship between assessment and learning. We have the tests we deserve because we are wont to reduce "assessment" to "testing" and to see testing as *separate* from learning—something you do expediently, once, after the teaching is over, to see how students did (usually for *other* people's benefit, not the performer's). Standardized multiple-choice tests thus represent an extreme version of an endemic problem: under these unspoken premises, it is inevitable that we come to rely on the most simple, efficient, and trivial tasks that can be reliably used. Saving time and labor becomes a paramount design virtue and causes every designer to rely on simple, quickly employed, and reusable (hence "secure") items.

The willingness by faculties to machine-score a local test or refer student answers to a key devised for quick marking during a faculty meeting reveals that teachers too, not just professional test makers, think that human judgment is an unnecessary extravagance in sound assessment. "But we teach so many students! What else can we do?" Why not see it the other way around? Why have the band director, debate coach, science fair judge, play director, and volleyball coach not succumbed to the same thoughtless or fatalistic expediency? Because they understand that the "test" of performance *is* the course, not something you do *after* it. Because they understand that both validity and reliability of judgment about complex performance depend upon many pieces of information gained over many performances. If our thoughtless assessment practices are going to change, we need to do more than simply replace traditional forms of "test" (multiple-choice) with new forms of "test" ("perfor-

mance" or "portfolio"); we need to change the fundamental relationship between tester and student.

Let me put this more forcefully, as a proposition that provides a rationale for the book: *tests are intrinsically prone to sacrifice validity to achieve reliability and to sacrifice the student's interests for the test maker's.* All testing involves compromise. Tasks are simplified and decontextualized for the sake of precision in scoring. Limits are placed on the student's access to resources. Standardization establishes a distance between tester and student: the student can neither adapt the question to personal style nor question the questioner (a right, one would think, in a modern society, and something we would encourage in a more personal educational relationship). Many of the *inherently* questionable practices in our "slender technique" (such as excessive secrecy and the inability of the student to ask questions or use apt resources, discussed in Chapters Three, Four, and Five) are of such long standing in testing that we do not see their potential for harm.

What is educationally vital is inherently at odds with efficient, indirect testing and unambiguous test items. As senior Educational Testing Service (ETS) researcher Norman Frederiksen has put it, using more formal measurement terms, "Most of the important problems one faces in real life are ill structured, as are all the really important social, political, and scientific problems in the world today. But ill-structured problems are not found in standardized achievement tests. . . . We need a much broader conception of what a test is if we are to use test information in improving educational outcomes."[3] And by an overreliance on tools that reduce thoughtful discernment of achievement to adding or subtracting points, we have also unwittingly caused a progressive illiteracy about assessment that is visible everywhere in American education.[4]

Whatever the original intentions and benefits of efficient and reliable testing, we thus perpetually engage in self-deception about its impact. As Dennie Wolf, Janet Bixby, John Glen, and Howard Gardner of Project Zero at the Harvard University Graduate School of Education noted recently, in a comprehensive discussion of new forms of assessment, "The irony of social inventions is that one-time innovations turn to habit."[5] The self-deception consists in thinking that our habit of testing has no impact: the very word

instrument as a synonym for *test* implies this illusory neutrality.[6] But a "test," no matter what the maker's intent, is not in fact a neutral instrument any more than tax tables are. When we assign value, we produce an impact: what gets measured gets noticed; what you test is what you get; what gets inspected gets respected. These and similar aphorisms point toward an inescapable cautionary lesson, particularly since tests inherently measure only what is easy to measure. (Wiser folks have called for the use of the most unobtrusive measures possible.[7])

The limits of all tests would be less of a problem if we were working sensibly and knowledgeably within those limits. John Dewey presciently warned us of the problem when standardized tests were first introduced: "The danger in the introduction of standardizing methods . . . is not in the tools themselves. It is in those who use them, lest they forget that it is only existing methods which can be measured."[8] What we must now face up to is that we have allowed a thoughtless proliferation of such tests in our educational world, without considering their limiting effects on pedagogy. As I will show, these ubiquitous "instruments" keep schools reliant on premodern psychological, moral, and epistemological assumptions about learners.

In our so-called culture of testing, "intelligence" is fixed. As a result, as Wolf and her colleagues have noted, "relative ranking matters more than accomplishment." The "easily quantifiable" is more significant than the "messy and complex." The "dominant image" of this culture is the "normal curve," in which the comparison of students against one another is viewed as more significant than the reporting of performance against standards.[9] In this testing culture, the use of one-shot events in June (instead of longitudinal assessment) is universal—despite the fact that the results are no longer useful to the performer (and that the reports from national tests are decipherable only by measurement experts).

Is it any wonder, then, that a fatalism pervades education? The tests we use result in a self-fulfilling prophecy about student ability—produced, in good measure, by the artifices of testing formats, schedules, and scoring mechanisms. Few educators really end up believing (no matter what a school mission statement says) that "all children can learn" to high standards, given that test results

rarely show dramatic change. (And they *won't* show such change, because of their design and our current patterns of use.) To develop an educational system based on the premise that "all children *will* learn," we need assessment systems that treat each student with greater respect, assume greater promise about gains (and seek to measure them), and offer more worthy tasks and helpful feedback than are provided in our current culture of one-shot, "secure" testing.

The Morality of Testing

This book offers a philosophy of assessment based upon a basic and decidedly more modern and unfatalistic principle: because the *student* is the primary client of all assessment, assessment should be designed to *improve* performance, not just monitor it. Any modern testing (including testing used for accountability) must ensure that the primary client for the information is well served.

But any philosophy of student assessment depends upon more than a value statement as to what assessment "should" be. A philosophy should lay out a coherent and thorough position describing how the theory and practice of testing and assessment should be best understood. Any philosophy of assessment must therefore provide a set of justified principles to serve as criteria for "testing the test." Presenting these principles is a primary aim of this book. How would such criteria differ from technical standards (those already established by psychometricians, for example)? The simple answer is that the tester's standards can tell us only when tests are sound or unsound, not when we should and should not test. To answer the latter question, we need to raise questions normally not raised in such discussions: When is testing apt, and when is it not? What should happen when the interests of test makers, test users, and test takers diverge? How should schools and districts establish safeguards against the harm of excessive testing and inadequate assessment? These are questions crying out for consideration.

These questions are pressing not merely because we are in the midst of unparalleled debate about testing and schooling. A logically prior question has never been adequately considered in the testing debates: Are the primary client's interests served by testing?

As I hope to show, a preponderance of testing (as opposed to assessment) is *never* in the student's interests, whether we use multiple-choice or performance-based tests. Because a test, by its design, is an artifice whose audience is an outsider, whose purpose is ranking, and whose methods are reductionist and insensitive.

Our habits run too deep for mere technical scrutiny. The questionable practices used in testing are not rational, no matter how ubiquitous; they cannot be understood on technical or policy grounds alone. Our "modern" testing systems are built upon an ancient human urge to perpetuate a "marking and classification" system, in Michel Foucault's words. (Stephen Jay Gould's history of intelligence testing in *The Mismeasure of Man* makes clear how easily ethnocentric assumptions infected the testing but were hidden by the methods.[10]) Our tests, as the chapter-opening quote from Foucault suggests, still reflect premodern views about the student's (unequal) relationship to the examiner (and hence premodern views about the student's rights). The "rituals" and "roles" of the examination are rooted in a medieval world of inquisitions and class distinctions. The examiner's "methods," then as now, prevent the student from questioning the tester about the questions, the methods, or the results—especially because of "secure" testing, a legacy of the secretive methods employed by medieval guilds. The use of such chicanery as "distracters" on tests and win-lose grading systems are among many practices that mark the testing relationship as morally imbalanced.

When our sole aim is to measure, the child is invariably treated as an object by *any* test, a theme to which I will repeatedly return. The educative role of genuine assessment is always at risk if by *test* we mean a process in which we insist upon *our* questions, *our* timing, and *our* imposed constraints on resources and prior access to the questions and in which we see *our* scoring and reporting needs as paramount. When we isolate the learner's "knowledge" from the learner's character in a test, we no longer feel an obligation to get to know the assessee well (or even to pursue the meaning of an answer). It is no longer obligatory to worry whether we have provided students with the opportunity to have their achievements thoroughly examined and documented; it is no longer obligatory to

BEST COPY AVAILABLE

construct opportunities for them to show what they can do (or, at the very least, to sign off on the results).

There is an *inescapable* moral dimension, in other words, to the assessment relationship—a dimension that we ignore. In school testing as we have always known it, that relationship is inherently tilted in favor of the tester. The tenor of the relationship has little to do with what *kind* of test we use and everything to do with the manner in which testing and other assessment methods treat the student, as we shall see in the chapters that follow. A philosophy of assessment in which student interests are viewed as primary would have us ask, What approaches to assessment are most respectful?

Respectful may seem like an odd word to use in talking about quizzes, tests, exams, grades, and the like, but it is the most apt word with which to initiate the rethinking of our deep-seated habits about testing. The assessor either respects or disrespects the student by the manner in which the relationship is conducted. It is respectful, for example, to be open with people about your intent and methods; a steady dose of secure tests must then be disrespectful and off-putting. It is respectful to allow people to explain themselves when we think that they have erred or when we do not understand their answer. Tests that provide no opportunity for students to supply a rationale for answers reveal that what they think and why they think it is unimportant. It is respectful to give people timely, accurate, and helpful feedback on their "effect," yet most tests provide the student with nothing more than a score—and often weeks later. It is respectful to give people ample opportunity to practice, refine, and master a task that we wish them to do; yet secure, one-shot tests prevent the efficacy that comes from cycles of model/practice/feedback/refine.

The assessment relationship can thus range from being one of complete mutual respect—through ongoing responsiveness, flexibility, patience, and a capacity for surprise in our questioning and follow-up—to one in which the student is only "tested" (and is thus treated as an object), through the imposition of tasks and procedures that provide the student (and teacher, in the case of externally designed tests) with no opportunity to enter the conversation.

The Epistemology of Testing

There are also troublesome epistemological questions that have never been properly addressed by most test designers, be they psychometricians or teachers. What really counts as evidence of "knowledge"? What is it that we want students to be able to do as a result of schooling? The de facto answer, if we examine tests, is that the student must recognize or plug in correct facts or principles to atomistic questions simplified of their natural messiness by the test maker. But is this a defensible view of what constitutes a knowledgeable student or a successful schooling? Clearly not, as we see quickly if we substitute the phrases "know-how" or "wisdom" for "knowledge."

What matters in education is understanding and the habits of mind that a student becomes disposed to use. Conventional testing cannot tell us, then, what we need to know, as I argue in Chapter Two—namely, whether the student is inclined to be thoughtful and able to be effective. Wolf, Bixby, Glen, and Gardner make the same point in quoting from William James: "Be sympathetic with the type of mind that cuts a poor figure in examination. It may be, in the long examination which life sets us, that it comes out in the end in better shape than the glib and ready reproducer, its passions being deeper, its purposes more worthy, its combining power less commonplace, and its total mental output consequently more important."[11] As Wolf and her colleagues argue, we are meant to understand that a thorough assessment requires describing long-term and multifaceted accomplishments. How else will we determine whether essential *habits* exist? (This is a validity question rarely addressed by measurement specialists.)

The problem is clearer still if we ask what counts as evidence that the student *understands* what was taught. A one-shot, secure test in which the student is required neither to produce a work-product nor to engage in discussion is unlikely to tell us whether the student has understanding or not. Correct answers can hide misunderstanding; incorrect answers, without an opportunity to explain oneself, can easily hide deeper insight. In traditional testing, the "number of items correct, not the overall quality of performance, determines the score. . . . It is as if the number of completed

sentences in an editorial mattered more than the overall power of the argument or the integrity of its perspective."[12] In Chapter Seven, I consider these issues from the perspective of measurement (what is validity?) and epistemology (what is the most vital meaning of "knowing" and "understanding"?).

The uniformity of test questions and the view that mastery means displaying common knowledge in prescribed forms also represent the persistence of premodern views. Prior to the scientific and democratic revolutions of the seventeenth and eighteenth centuries, it was assumed that knowledge was uniform, stable, and "theoretical," attained through didactic teaching, private reflection, and reading and writing (as opposed to through argument, action, discovery, or experimentation). Control over "truth" was the student's paramount obligation; intellectual autonomy and meaning-making—and with them, a *diversity* of understandings—would have been viewed as heresy. Thus to "test" the student was, and is, a practice of determining whether the student has mastered what is orthodox. The absence of opportunities to display one's understanding in one's own way derives from the "slender technique" of the medieval examination.

A sign that assessment reform is fighting ancient myths about learning and modern impatience with complex assessment can be seen in our deeper ignorance of the venerable Taxonomy. Benjamin Bloom, who developed the Taxonomy, and his colleagues were quite clear that a synthesizing understanding was displayed through diverse and even unpredictable student action: "The student should . . . be made to feel that the product of his efforts need not conform to the views of the instructor, or some other authority . . . [and] have considerable freedom of activity . . . [including] freedom to determine the materials or other elements that go into the final product."[13]

Synthesis is thus *inherently resistant* to testing by multiple-choice or other methods that assume uniform, correct answers, because it requires the student to fashion a "production of a unique communication" that "bears the stamp of the person."[14] Not only is diversity of response to be expected; the correct answer may not even be specifiable: "Synthesis is a type of divergent thinking: it is unlikely that the right solution can be fixed in advance." Higher-

order assessment will therefore almost always be judgment-based: "Each student may provide a unique response to the questions or problems posed. It is the task of the evaluator to determine the merits of the responses." (We must develop apt scoring criteria and standards that reward diverse excellence, in other words.) Nor will standardized testing conditions likely be appropriate: "Problems should be as close as possible to the situation in which a scholar/artist/engineer etc. attacks a problem. The time allowed, conditions of work etc. should be *as far from the typical controlled exam situation as possible*. . . . It is obvious that the student must have considerable freedom in defining the task for himself/herself, or in re-defining the problem or task."[15]

How have we allowed this commonsense point of view to be repeatedly lost? One answer is suggested by Foucault's observation that there is "a whole type of power" in testing. That we glibly talk about "instruments" and "items" is only one indication that the *format* of one type of test has greatly influenced our thinking about what counts as knowledge and as evidence of mastery. Large test companies have made millions of dollars making and selling tests to school systems, claiming that the multiple-choice items test for higher-order thinking and achievement. Whether we consider Bloom's or anyone else's views of higher-order thinking, this is a dubious claim at best: there is plenty of evidence to show that these test-company claims simply do not hold up under disinterested scrutiny.[16] (Why do we allow testers to be judge, jury, and executioner in their own case—a situation without precedent in American consumer affairs?)

The ironic and unfortunate result is that teachers have come to resist formal evaluation of all kinds, given the intellectual sterility and rigidity of most generic, indirect, and external testing systems. Because of that resistance, local assessment practices are increasingly unable to withstand technical scrutiny: teacher tests are rarely valid and reliable, and "assessment" is reduced to averaging scores on tests and homework. We would see this all more clearly if we grasped the implications of the idea that assessment should *improve* performance, not just *audit* it. As long as assessing is just seen as testing, and "testing" amounts only to putting numbers on papers and letters on transcripts, the full harm of our practices to

the learner goes unseen. But harm there is. Learning *cannot* take place without criterion-referenced assessment, no matter how good the teaching. Successful learning depends upon adjustment in response to feedback; no task worth mastering can be done right on the first try. Effective adjustment depends upon accurate self-assessment; good self-assessment depends upon the assessor supplying excellent feedback—useful as judged by the performer—and prior insight into the standards and criteria of assessment. (See Chapter Six, which addresses feedback, and Chapter Three, which explores the debilitating impact of a world of testing that precludes the *possibility* of feedback—namely, the use of secrecy before, during, and after the test.)

Apt feedback depends upon the sharp senses of a judge who knows how to assess current performance in light of standards and criteria—that is, the hoped-for results—while remaining mindful of the character of the performer and the context of the assessment. Generic assessment is a contradiction in terms, in other words. Good assessment always involves tact, in the older sense of that word, and tact is always threatened by testing (a notion developed in Chapter Four). Effective feedback—namely, feedback that helps the learner improve—is *impossible* without such tact.

The message of William Barrett's book on philosophy of a few years back, *The Illusion of Technique*, is thus an apt one for summarizing testing in schools: we have made generic what must be relational and situational; we have made what is inherently murky artificially precise.[17] By making a practice that ought to be a focal point for learning into an after-the-fact checkup not worth dwelling upon, we have unwittingly caused the most vital "organ" of pedagogy and genuine intellectual assessment to begin to atrophy in our teachers: human judgment. How? By reducing assessment to testing. This book is about why schools must fight to make assessment of student work primary and testing subservient to a sound assessment strategy.

Assessment Versus Testing

The distinction between an *assessment* and a *test*, made often in the previous pages, is not merely political or semantic (in that derogatory sense of hairsplitting). An assessment is a comprehensive, multifaceted analysis of performance; it must be judgment-based and personal. As Lee Cronbach, Stanford University professor and the dean of American psychometricians, put it over thirty years ago, assessment "involves the use of a variety of techniques, has a primary reliance on observations (of performance), and involves an integration of (diverse) information in a summary judgment." As distinct from "psychometric measurement" (or "testing"), assessment is "a form of clinical analysis and prediction of performance."[18] An educational test, by contrast, is an "instrument," a measuring device. We construct an event to yield a measurement. As Frederiksen puts it, "A test may be thought of as any standardized procedure for eliciting the kind of behavior we want to observe and measure."[19]

But the *meaning* of the measurement requires assessment. Assessment done properly should *begin* conversations about performance, not end them. Even the father of intelligence testing understood that one test was not a sufficient indicator of a person's capacities. Alfred Binet was anxious to see his tests used as part of a thoughtful assessment process. He warned his readers, for example, that, "notwithstanding appearances, [the intelligence tests are] not an automatic method comparable to a weighing machine in a railroad station. . . . The results of our examinations have no value if deprived of all comment; they need to be interpreted."[20] And he warned his readers in the last version of the IQ tests that "a particular test isolated from the rest is of little value; . . . that which gives a demonstrative force is a group of tests. This may seem to be a truth so trivial as to be scarcely worth the trouble of expressing it. On the contrary, it is a profound truth. . . . One test signifies nothing, let us emphatically repeat, but five or six tests signify something. And that is so true that one might almost say, 'It matters very little what the tests are so long as they are numerous.' "[21] This is a warning rarely heeded by teachers. Many students have paid dearly for their teachers' haste and impatience with testing and grading—haste bred by school schedules that demand 128 final grades by the Monday after the Friday exam. (Measurement specialists forget that lots of similar items, using only one format, do not count as "many tests.")

It thus makes sense to distinguish between performance *tests*

and performance *assessments* in just the ways suggested by Cronbach, despite the fact that many advocates of performance testing refer to their tests as performance assessments. A performance test is meant to yield a score, albeit one that some believe to be more valid than the score yielded by an indirect test. A performance *assessment*, on the other hand, is meant to yield a more comprehensive judgment about the meaning of this score and performance in general, viewed in various ways. The assessor may or may not use "tests"—of a direct or an indirect kind—as part of any assessment.

We need to keep in mind, therefore, that the central question is not whether we should use one kind of test or another but what role testing should play in assessment. Criticism of testing should be understood as concern for the harm that occurs when (judgment-based) assessment is reduced to (mechanical) testing and conclusions based on test data alone.

The etymology of the word *assess* alerts us to this clinical—that is, client-centered—act. *Assess* is a form of the Latin verb *assidere*, to "sit with." In an assessment, one "sits with" the learner. It is something we do *with* and *for* the student, not something we do *to* the student. The person who "sits with you" is someone who "assigns value"—the "assessor" (hence the earliest and still current meaning of the word, which relates to tax assessors). But interestingly enough, there is an intriguing alternative meaning to that word, as we discover in *The Oxford English Dictionary:* this person who "sits beside" is one who "shares another's rank or dignity" and who is "skilled to advise on technical points."

Technical soundness in student testing is not enough, in other words. In an assessment, we are meant to be the student's moral equal. (At the very least, as Peter Elbow, long-time researcher and thinker on writing and teaching, argued in discussing competency-based teaching, the teacher should move from being the student's "adversary" to "ally."[22]) Such a "sitting with" suggests that the assessor has an obligation to go the extra mile in determining what the student knows and can do. The assessor must be more tactful, respectful, and responsive than the giver of tests—more like "a mother and a manager," in British researcher John Raven's phrase, "than an imperious judge."[23] One might go so far as to say that the assessor (as opposed to the tester) must ensure that a student's strengths have been found and highlighted (irrespective of what weaknesses one also documents). This is precisely what the Department of Labor's SCANS (Secretary's Commission on Achieving Necessary Skills) report on education and the workplace called for in saying that students ought to leave high school with a "résumé," not a transcript.[24]

A *test*, however, suggests something very different. It is an evaluation procedure in which responsiveness to individual test takers and contexts and the role of human judgment are deliberately minimized, if not eliminated. This is intended as an observation, not a criticism. There are well-known virtues to standardizing procedure and minimizing bias, drift, and other forms of error in judgment. Most tests accomplish this mechanization of scoring by taking complex performances and dividing them into discrete, independent tasks that minimize the ambiguity of the result. (We can take off points and count up scores easily, in other words.) As a result, most tests tend to be "indirect" (and thereby inauthentic) ways of evaluating performance, because tests must simplify each task in order to make the items and answers unambiguous and independent of one another. As we shall see, we have paid a price for this inauthenticity, irrespective of the indisputable precision gained by using indirect measures.

An inherent tendency toward simplification is not the only danger in testing. For a variety of policy-related and historical reasons, testing in this country has become generic as well, in the sense of being linked neither to a particular curriculum nor to realistically complex problems and their natural settings. We live in a schizophrenic world of shared national expectations for schools but diverse local cultures and curricula. We have defined *accountability* as comparability on common measures, despite the fact that accountability is not dependent upon tests and is better done at the local level through responsiveness to clients (see Chapter Eight). Because of our diversity, our ability to do accurate comparisons across classrooms, schools, and districts is dependent on questions that are not highly contextual or articulated with local curricula.

But a generic test of understanding is a contradiction in terms. We want to know about *this* child, in *this* setting, in relation to *this* curriculum. It is for this reason that Howard Gardner argues

that "what distinguishes assessment from testing is the former's favoring of techniques which elicit information in the course of ordinary performance, and its general uneasiness with the use of formal instruments administered in a neutral decontextualized setting."[25]

As a developmental psychologist, Gardner no doubt was influenced in making this distinction by the work of Jean Piaget in the use of clinical interviews and observations. It was Piaget who first cautioned about the limits of testing in assessment, troubled by the artificiality of forced responses to narrowly defined questions and tasks. The test is modeled on the scientific experiment, he reminds us. Given that, the process must be both controlled and replicable, hence "standardized."[26] One variable—the particular skill or fact—is isolated by the test question even though this isolation makes the test unrealistically neat and clean ("well structured," in measurement jargon).

The clinical assessment, by contrast, can never be rigidly standardized, because the interviewer must be free to diverge from a protocol or strategy of questioning in order to follow up on or better elicit the student's most revealing acts and comments. The specific questions and physical tasks that are used by the assessor are meant to serve as mere prompts to revealing and spontaneous action; we vary the questioning, as necessary, if we feel that the answers are somehow not genuine or revealing. The aim is to use both anticipatable and *unanticipatable* responses to assess the student's actions and comments. (Piaget repeatedly warned about the likelihood of inauthentic responses by the child, whether due to unwittingly suggestive questioning by the adult, boredom on the child's part, or a desire by the child to tell the questioner what the child thinks the questioner wants to hear. Thus varying the questions may be essential to procure authentic responses.[27])

The assessor tries to ferret out all of what the student knows and can do by various means. The tester, on the other hand, demands of the student specific responses to fixed questions of the tester's choosing. The student does not have the freedom (the right, really) to say on tests, "Wait! Let me reframe the question or clarify my answer!" The format of the test may be modeled on modern scientific processes, but the philosophical assumptions that permit

the student to be treated as an object of the tester-experimenter are premodern.

At the very least, assessment requires that we come to know the student in action. Assessment requires a "multidimensional attempt to observe and to judge the individual learner in action" using "careful judgment," as the faculty of Alverno College in Milwaukee describe the philosophy and practice of two decades of their renowned outcomes-based liberal arts program.[28] In their materials, they stress repeatedly that the purpose of assessment is to assist and inform the learner. That is why so much of their formal assessment involves a formal self-assessment process as the cornerstone of later achievement and autonomy (a process that I discuss further in Chapter Two).

The Alverno understanding also gives us an insight into why assessment and performance have been historically linked. Self-assessment makes sense only in a world of testing where the criterion performance is known and serves as a clear and focusing goal. (What sense would there be in self-assessment in reference to indirect, one-shot, simplistic tests?) While a test in which the student responds to prefashioned answers tells us what the student "knows," it does not tell us whether the student is on the road to using knowledge wisely or effectively. As the faculty at Alverno put it, "Narrow, one-dimensional probes into a student's mines of stored information do not begin to get at how she learns or what she can do."[29] At Alverno, each student's progress and graduation are dependent upon mastering an increasingly complex set of performance tasks that simulate a variety of professional-world challenges. What assessment entails is insight into academic personality—intellectual character.

We now move from a women's college to wartime recruitment of spies. The idea of using many different tests to conduct an assessment of a person's character and the use of authentic simulations lay at the heart of Office of Strategic Services (OSS) recruitment procedures during World War II. The team of psychologists and psychiatrists charged with identifying the best candidates for work behind enemy lines developed what they called a "multiform, organismic (i.e., holistic) system of assessment: 'multiform' because it consists of a *rather large number of procedures based on different*

principles, and 'organismic' (or 'holistic') because it utilized the data obtained through these procedures for attempting to arrive at a picture of the personality as a whole."[50]

Whether or not the term *assessment* first gained wide usage through the OSS "assessment of men" (as Russell Edgerton, president of the American Association for Higher Education, has claimed[51]), we do know that the research team had an explicit and compelling interest in distinguishing between what they called "elementalistic" testing and "holistic" assessing. "Elementalistic" testing is testing as we know it: complex performances are turned into a small set of abstracted "operations," each of which can be objectively scored. The OSS perceived such testing as "scientific" but "abstract and unrealistic."

The realism that the OSS felt to be essential was introduced by a clever series of simulations—what OSS assessors called "situational tests"—designed to replicate not only the challenges but also the conditions the recruits were likely to face. (This was necessary because they routinely found a lack of correlation between certain paper-and-pencil tests and actual performance, particularly where social interaction was central to the task.) The tasks are familiar to many of us now through such programs as Outward Bound and Project Adventure: tackling a ropes course, negotiating a maze, staying in fictitious character throughout an interrogation, bringing a group task to completion (despite planted confederates of the staff, who do their best to screw up the mission and question the recruit's authority), memorizing a map for later use, and so on.

The material gained from clinical interviews and the staff's overall judgment could and did override any mechanical decision made as a result of each test and total scores: "It was one of the noteworthy features of the OSS assessment system that it recognized explicitly the necessity of relating all observations to each other, not in a mechanical way, but by an interpretive process." It was this viewpoint that made the team rely heavily on "additional procedures" beyond traditional tests. They found the use of "autobiography, interviews, situational tests, psychodramas, and projection tests" indispensable in developing a clear picture of the whole person and his or her likely performance.[52]

BEST COPY AVAILABLE

Testing: Standards and Laws

There is nothing radical or "soft" about an emphasis on the whole person and the need to use judgment in assessing the meaning of scores. The American Psychological Association/National Council on Measurement in Education/American Educational Research Association Standards for Educational and Psychological Testing (hereafter referred to as the APA Standards) are unequivocal on this matter. Standard 8.12 says that "in elementary and secondary education, a decision that will have a major impact on a test taker should not automatically be made on the basis of a single test score. Other relevant information for the decision should also be taken into account by the professionals making the decision."[53] There are also ten standards subsumed under Standard 16 ("Protecting the Rights of Test Takers") that have to do with such matters as informed consent and the right to privacy about scores.

But where under Standard 16 is a discussion on students' right to minimal test security? What about the right of access to test papers after they have been scored? Where are the standards that provide students with the right and realistic opportunity to challenge a score or the aptness of the questions? (Incredibly, the final four substandards under Standard 16 discuss the rights of due process when a test score is thought to be fraudulent! Thus one's "rights" only come into play when one is accused of criminal behavior, not as a consumer or citizen entitled to knowledge concerning decisions affecting one's fundamental interests.) During the 1970s, when consumer rights came to the fore, the APA Ethical Principles in the Conduct of Research with Human Participants did call for specific protection on some of these points; Walt Haney and George Madaus, Boston College measurement experts, quote: "Persons examined have the right to know results, the interpretations made, and, where appropriate, the original data on which the final judgments were made." But in the 1981 and 1985 versions of the Ethical Principles, the section was deleted. (And a decade-old suit by the test companies to keep tests secure after administration still languishes in the courts, in a challenge to New York's law on open testing.) Haney and Madaus also note, in describing the changes in the APA Standards, that "strong vested interests within

BEST COPY AVAILABLE

the APA demanded changes that favored their constituents at the expense of the test taker. . . . The profession has yet to come to grips with how the Standards can work fairly and equitably for developer, user, and test taker without an enforcement mechanism other than the courts."[34]

Indeed, it has been in the courtroom that some of the more sacred testing practices have been questioned. Unbeknownst to most teachers, there are a variety of legal precedents in which the claims of testing companies and test-using institutions about the validity of scores have been successfully challenged. *Debra P. v. Turlington* (644F.2d.397[1981], for example, a case involving a failing grade on a state minimum-competency test, yielded a ruling that a test had to *match what the students were in fact taught* for a score to be valid. (For brief reviews of the legal history of testing, see Cronbach, 1990, and the Introduction in Berk, 1986.)[35]

One reason that educators might not be aware of this legal history is that many cases have centered on employment decisions in the adult workplace. Despite the workplace focus, however, the legal precedents might have an effect on the testing debate in schools if teachers, students, and parents were more aware of their rights, because many of those precedents bear on the current interest in performance testing.

Most of the recent court cases grew out of Title VII of the 1964 Equal Employment Opportunity Act and the establishment of the Equal Employment Opportunity Commission (EEOC); they have to do with charges that testing procedures inappropriately discriminate. But what may surprise some readers is that many cases involve successful challenges to the inappropriate content of multiple-choice standardized tests. In those cases, the plaintiff's argument often rested on the claim that the test's content had little to do with the proposed job or performance on it; the companies and/or test makers were cited for the failure to validate their tests through a careful job analysis—that is, to validate the proposed test tasks against the actual abilities, knowledge, and attitudes required by a job. (In one case, not even a job description was viewed as adequate unto itself; nor were supervisors' ratings deemed adequate.) In fact, both the EEOC Guidelines and the APA Standards

referred to earlier specify that a thorough job analysis is required in employment- or licensure-focused testing.

Thus in *Fire-fighters Institute for Racial Equality* v. *The City of St. Louis* (616F.2d350[1980]), minority firefighters twice successfully sued on the basis that the skills tested were too dissimilar to those used by fire captains on the job. In another case that went to the U.S. Supreme Court (*Albemarle Paper Co.* v. *Moody*; 422US405[1975]) the court struck down the use of an employer's test because no job analysis had been conducted to show that the skills cited as essential by the company in its own validation study were in fact needed for the jobs listed.

I am not a lawyer. My aim is not to dissect cases but to raise an important question suggested by a cursory look at the legal history: Why have these standards for and concerns with job analysis rarely been applied to educational testing? They are not completely absent, fortunately. In fact, there is a lengthy and important history in testing and curriculum at the *adult* level concerning the need to start with task or role analysis.[36] There are also sweeping changes underway in licensure to make the process more performance-based at the national level.[37] And the role analysis on which such design is based is at the heart of any competency-based education, as many writers have pointed out.[38]

A demand for authenticity is therefore more than just a demand for face validity, despite what some caustic traditionalists have claimed. It is a demand for tests built upon intellectual job analyses. Since the validity of educational tests is now almost universally achieved merely through correlation with other test scores or grades in college, this kind of careful "job analysis" (that is, correlation with the ultimate criterion performances) is rarely done. And the "job" of performing well in professional or civic life has little or nothing to do with the kinds of tasks and content-related questions found on most commercial and teacher-constructed tests. Indeed, this was a central element of the argument made by David McClelland, a Harvard University psychologist, in his renowned critique of testing twenty years ago (discussed in Chapter Seven).[39]

Nor should those who seek to replace multiple-choice testing with performance testing rest easy on the legal front. The courts have consistently found for the plaintiffs in situations where there

were no formal guidelines or procedures to ensure interrater reliability in judgment-based forms of testing, for example. (On that ground alone, most of our schools would be found guilty of violating student rights.)

Sound assessments do not differ from tests simply because the former are more complex than the latter. The questions of rights and responsibilities are crucial: in a proper assessment, we put the student's rights first; in an imposed test, with no oversight, we put the tester's interests (not rights, mind you) first. The "assessment" of reading by a portfolio and interview thus easily turns into an overly emphasized and perhaps unreliable test if our primary purpose is to report a single, efficiently gained score to compare with the scores of others. The "test" of Outward Bound or putting on a play is closer to what we mean by true assessment—educative on many levels about both the student and the criterion situation. These challenges are very different in tone and consequence from those "performance assessments" that still seek a single score—just a "better" one.

The Standards of Assessment

None of this should be heard—as I know it often is—as the sanctioning of assessment understood as coddling or merely praising, a flinching from telling the student where the student truly stands. I do not support the efforts of those reformers who see the employing of rigorous criteria and standards as somehow anti-assessment and anti-student.[40] As I have repeatedly argued (in consulting, in workshops, in the CLASS video series on how to undertake assessment reform, and in print), for local work to be both useful and credible, faculties must "benchmark" their grading and work to develop more criterion-referenced procedures and better interrater reliability in their grading.[41] This is especially important in light of the well-known lack of score reliability to be found in the judgments of naive and untrained assessors of complex performance. Vermont's portfolio system, for example, has already run into this problem.

Nor are the arguments for standard-referenced assessment related primarily to school accountability concerns. As I shall argue in Chapters Five and Eight, the *learner's* performance can be improved only when it is measured against "progress," not "growth." And measuring progress means measuring "backward" from the "destination"—that is, the standards and criteria exemplified in models of excellent work. Mere *change* in the student's performance provides no insight into whether that change is likely to lead to excellent adult performance. To both help the learner and inform the parent and the community, the work of assessment must be done in reference to clear and apt criteria and standards.

There is thus no simple, either-or choice about what to test, how to test, and how to assess using tests. Instead, there are unending dilemmas. The challenge is to develop a picture of the whole person using a variety of evidence, including tests, and to do so in the most feasible and helpful (to the student) way. One of the many dilemmas is that testing so easily becomes the entire assessment system instead of being a facet of it; testing is, after all, quicker, cheaper, neater, and cleaner. (This and other dilemmas are mapped out more thoroughly in Chapter Two.) We can see a kind of Gresham's law in education to be eternally resisted: "Efficient tests tend to drive out less efficient tests, leaving many important abilities untested—and untaught."[42] It may well be in the learner's interest to be tested, and the student may well learn something from being tested, but a preponderance of mere testing puts the learner's intellectual interests at risk. What is at stake, then, is whether we will insist on an assessment process in which testing practices respect the student's long-term interests.

This book should be understood, then, as a call to develop policies and principles for the testing of tests in a way that better protects students. There are more than technical and strategic questions at stake in testing. The fundamental quandary is how we can best employ testing as one facet of an overall assessment strategy that serves all "customers for information" but sees the student "customer" as primary. We need to consider the intellectual and, yes, moral consequences of testing—especially the harm inflicted by the pervasive secrecy that dominates testing, the unfairness of scoring systems with built-in disincentives, the ethics of using test "distracters," and the tactlessness involved in using answer keys that admit no ex post facto correction on the basis of an apt but unanticipated answer.

Yet there is undoubtedly a risk in drawing upon the language and logic of morality (and epistemology) in discussing student assessment and testing, because to do so may appear either naive or unhelpful to those who must address the pressing problems of testing and school reform. Nonetheless, it is essential that we take the risk. *Any use of power and its limits involves moral questions and moral consequences, and testers have extraordinary and often unilateral power.* Thus nothing could be more practical than to develop procedures whereby the logically prior questions about the purposes and effects of assessment and testing are thoroughly considered and codified into intelligent policies and practices.

Every test, every grade affects the learner. Every dull test—no matter how technically sound—affects the learner's future initiative and engagement. No, even saying it this way does not do justice to the consequences of our testing practices: every test *teaches* the student. It teaches the student what kinds of questions educators value, it teaches the student what kind of answers we value (correct merely? justified? chosen from our list? constructed by the student?), it teaches the student about intellectual relationship, and it teaches the student how much we respect student thinking.

Toward an Assessment Bill of Rights

We might, then, think of sound assessment and the appropriate use of tests along the lines of the modern judiciary or our own Bill of Rights. Due process is essential in good assessment, just as it is in legal proceedings. What counts as evidence must be acceptable not only to disinterested judges but also to such interested parties as the student and the student's teachers. The student/teacher should have not only the right of appeal but the right to question the questioner (to "cross-examine") and to present other kinds of evidence—in short, to make one's case. (That these rights sound fantastic and impossible, though they are routinely granted in the U.S. higher-education system and the K-12 systems of other countries, shows how far we have come in valuing efficiency over effectiveness, "measurement" over education.)

Cronbach makes explicit what is usually implicit in this regard: "The tester enters into a contract with the person tested. In former days the understanding was left vague. The tester is now expected to be frank and explicit. . . ." Cronbach offers some general guidelines about problematic practices ("Scores that will have an important effect on the person's future should be reported in understandable form," "A procedure for challenging a test report should be available and the test taker should be informed of the procedure," and so on) but warns that "such principles are not easily put into practice."[43]

But demanding that the tester be more "frank and explicit" is not enough. The power is still so unequally distributed and the methods used by the test maker are so arcane that the student (and the teacher, when the tests are external) has no real opportunity to understand the proceedings. As in the legal model, there ought to be not only an appeals process for testing and assessment decisions (as there often is in special education decisions, for example) but also "cross-examination" of test makers as to the compromises and limitations of the test design and the meaningfulness of the scores.

Local educators need greater guidance and more helpful decision-making structures as they attempt to balance the inherently competing interests of test giver and test taker. It is *never* in the student's interests for test questions and answer sheets to remain secure not only before but after administration, yet test companies and teachers have thus far successfully lobbied to retain the right to protect their product and save themselves time and money.

Principles for the conduct of assessment that put the student's rights on a moral par with the technical needs of psychometricians and the policy needs of school-board and community members are long overdue. I know of only a handful of school districts that have testing and assessment policies. And policy makers, educators, classroom teachers, and district administrators need significant help in "testing" the tests—help that takes the typical complaints beyond the merely personal to the principled.

More formal and powerful help could perhaps come from a genuinely disinterested national organization set up to review tests and test-maker claims, modeled along the lines of Underwriters' Laboratory or the Federal Trade Commission—an idea offered by Haney and Madaus and others as a more appropriate venue than the de facto one of courts of law: "[We call for] the establishment of an

independent auditing agency that would, without vested interest, evaluate tests and testing programs that profoundly affect the lives of examinees."[44] Legislation may well be part of the solution. In a recent article on what should be done at the federal level to prevent testing-related abuses, Michel Feuer, Kathleen Fulton, and Patricia Morrison of the Congressional Office of Technology Assessment made a similar pitch: Congress might "focus on various proposals to certify, regulate, oversee, or audit tests," including the establishment of an oversight agency.[45]

Feuer, Fulton, and Morrison also argue Congress should "require or encourage school districts to develop and publish a testing policy."[46] Each school district ought to, at the very least, state the permissible uses of such morally problematic practices as test security, scoring work on a curve, the use of nonarticulated and generic tests, the failure to require consistent grading among teachers for the same work, and so on—practices discussed throughout this book. These policies would do more than state the values of the institution; they would provide a local procedure for ensuring that assessment practices of both students and educators were publicly scrutinized, discussed, justified, and improved. Regional service agencies and federally funded educational labs and centers might also be asked to serve as a clearinghouse for policy statements and for guidelines in their formulation.

Here is a sample set of guiding principles from the New Zealand Department of Education:[47]

Principles of Assessment for Better Learning

1. The interests of the students shall be paramount. Assessment shall be planned and implemented in ways which maximize benefits for students, while minimizing any negative effects on them.
2. The primary purpose of assessment shall be to provide information which can be used to identify strengths and to guide improvement. In other words, it should suggest actions which may be taken to improve the educational development of students and the quality of educational programmes.
3. Assessment information should not be used for judgmental or political purposes if such use would be likely to cause harm to students or to the effectiveness of teachers or schools.
4. Every effort should be made to ensure that assessment and evaluation procedures are fair to all.
5. Community involvement is essential to the credibility and impact of assessment and evaluation processes. All parties with a direct interest should have an opportunity to contribute fully. Self-assessment is the appropriate starting point.
6. Careful consideration should be given to the motivational effects of assessment and evaluation practices.
7. In the assessment of intellectual outcomes, substantial attention should be devoted to more sophisticated skills such as understanding of principles, applying skill and knowledge to new tasks, and investigating, analyzing, and discussing complex issues and problems.
8. Emphasis should be given to identifying and reporting educational progress and growth, rather than to comparisons of individuals or schools.
9. The choices made in reporting assessment information largely determine the benefit or harm resulting from the information. For this reason, the selection, presentation, and distribution of information must be controlled by the principles outlined previously.

My own view is that, while these principles represent an important step toward protecting the student, they are too diffuse as stated. I would prefer that school systems develop an Assessment Bill of Rights to protect the inherently vulnerable student from the harms that testing easily leads to. It would be supported by explicit

audit or oversight policies to ensure that the rights were protected. Here is my rough draft of such a set of rights:

Assessment Bill of Rights

All students are entitled to the following:

1. Worthwhile (engaging, educative, and "authentic") intellectual problems that are validated against worthy "real-world" intellectual problems, roles, and situations
2. Clear, apt, published, and consistently applied teacher criteria in grading work and published models of excellent work that exemplifies standards
3. Minimal secrecy in testing and grading
4. Ample opportunities to produce work that they can be proud of (thus, ample opportunity in the curriculum and instruction to monitor, self-assess, and self-correct their work)
5. Assessment, not just tests: multiple and varied opportunities to display and document their achievement, and options in tests that allow them to play to their strengths
6. The freedom, climate, and oversight policies necessary to question grades and test practices without fear of retribution
7. Forms of testing that allow timely opportunities for students to explain or justify answers marked as wrong but that they believe to be apt or correct
8. Genuine feedback: usable information on their strengths and weaknesses and an accurate assessment of their long-term progress toward a set of exit-level standards framed in terms of essential tasks
9. Scoring/grading policies that provide incentives and opportunities for improving performance and seeing progress against exit-level and real-world standards

I am sorry to report that the idea of an Assessment Bill of Rights has been attacked by more than a few teachers when I have offered it in workshops. Some have actually angrily called for a *prior* list of student responsibilities (though I do not recall such a list in our Constitution). Perhaps nothing better illustrates why

these rights are deserving of formal protection, given the uneven balance of moral power in both the testing and teaching relationships as traditionally defined. The implicit hypocrisy in the position of these teachers is easily made explicit when one asks them whether they would be willing to endure a professional performance appraisal conducted under the same conditions as student testing.

Assessment, to be educative and fair, needs to be more a matter of principle and less a matter of good intentions, mere habit, or personality. In the chapters that follow, I explore some of the principles that underlie real and ideal testing and assessment—and the dilemmas that make it a profound mistake to replace sound assessment principles with unthinking procedures or answer keys requiring no judgment.

Notes

1. M. Foucault, *Discipline and Punish* (New York: Vintage Books, 1979), pp. 184–185. Translation Copyright © 1977 by Alan Sheridan.
2. D. P. Resnick and L. B. Resnick, "Standards, Curriculum, and Performance: A Historical and Comparative Perspective," *Educational Researcher* 14 (1985): 5–21.
3. N. Frederiksen, "The Real Test Bias," *American Psychologist* 39 (1984): 193–202, p. 199.
4. See, for example, R. Stiggins, "Assessment Literacy," *Phi Delta Kappan* 72 (1991): 534–539.
5. D. Wolf, J. Bixby, J. Glen III, and H. Gardner, "To Use Their Minds Well: Investigating New Forms of Student Assessment," in G. Grant, ed., *Review of Research in Education* (Washington, D.C.: American Educational Research Association, 1991), p. 31.
6. See H. Berlak and others, *Toward a New Science of Educational Testing and Assessment* (New York: State University of New York Press, 1992), p. 182ff.
7. See, for example, P. Terenzini, "The Case for Unobtrusive Measures," in Educational Testing Service, ed., *Assessing the Outcomes of Higher Education*, Proceedings of the 1986 ETS

BEST COPY AVAILABLE

Invitational Conference (Princeton, N.J.: Educational Testing Service, 1987).

8. J. Dewey, "Current Tendencies in Education," in J. A. Boydston, ed., *The Middle Works of John Dewey: 1899-1924* (Carbondale: Southern Illinois University Press, [1917] 1985), p. 119.
9. Wolf, Bixby, Glen, and Gardner, "To Use Their Minds Well," pp. 43-44.
10. S. J. Gould, *The Mismeasure of Man* (New York: W. W. Norton, 1981).
11. As quoted by Wolf, Bixby, Glen, and Gardner, "To Use Their Minds Well," p. 51.
12. Ibid., p. 47.
13. B. S. Bloom, ed., *Taxonomy of Educational Objectives*, Vol. 1: *Cognitive Domain* (White Plains, N.Y.: Longman, 1956), p. 173.
14. Ibid., p. 175. Compare B. S. Bloom, G. Madaus, and J. T. Hastings, *Evaluation to Improve Learning* (New York: McGraw-Hill, 1981), pp. 52-56.
15. Bloom, Madaus, and Hastings, *Evaluation to Improve Learning*, pp. 265, 268 (emphasis added).
16. See G. Madaus and others, *From Gatekeeper to Gateway: Transforming Testing in America* (Chestnut Hill, Mass.: National Commission on Testing and Public Policy, Boston College, 1990), and Frederiksen, "The Real Test Bias," for example.
17. W. Barrett, *The Illusion of Technique: A Search for Meaning in a Technological Civilization* (New York: Anchor Books, 1978).
18. L. J. Cronbach, *Essentials of Psychological Testing*, 2nd ed. (New York: HarperCollins, 1960), p. 582.
19. Frederiksen, "The Real Test Bias," p. 199.
20. A. Binet, and T. Simon, "The Development of Intelligence in the Child," in *The Development of Intelligence in Children* (Salem, N.H.: Ayer, [1908] 1983), p. 239.
21. A. Binet and T. Simon, "New Investigation upon the Measure of the Intellectual Level Among School Children," in *The Development of Intelligence in Children* (Salem, N.H.: Ayer, [1911] 1983), p. 329.

22. P. Elbow, "Trying to Teach While Thinking About the End," in G. Grant and Associates, *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education* (San Francisco: Jossey-Bass, 1979). See also P. Elbow, "One-to-One Faculty Development," in J. F. Noonan (ed.), *Learning About Teaching*. New Directions for Teaching and Learning, no. 4. San Francisco: Jossey-Bass, 1980.
23. J. Raven, "A Model of Competence, Motivation, and Behavior, and a Paradigm for Assessment," in Berlak and others, *Toward a New Science of Educational Testing and Assessment*, p. 100.
24. Department of Labor, *What Work Requires of Schools: A SCANS Report for America 2000* (Washington, D.C.: U.S. Government Printing Office, 1991).
25. H. Gardner, "Assessment in Context: The Alternative to Standardized Testing," in B. Gifford, ed., *Report to the Commission on Testing and Public Policy*, (Boston: Kluwer Academic Press, 1989), p. 90.
26. Critics of the multiple-choice test often do not grasp the fact that in *all* testing we "standardize" the conditions of test administration. We hold those conditions constant and we isolate knowledge to obtain a fixed answer to our specific question as the means of isolating a specific aspect of achievement with precision.
27. Whether developmental schemes such as Piaget's and Kohlberg's should be viewed as implicit systems of the "evaluation" of intellectual and moral behavior is an interesting and nettlesome question. In other words, should we say that empirical descriptions of growth constitute intellectual/moral progress, so that a "higher" score is "better" performance? Kohlberg thought so; Piaget did not (though Kolhberg thought he did). Gilligan, of course, thinks both schemes were empirically and conceptually flawed by the absence of adequate data and analysis of girls' moral and intellectual experience. See C. Gilligan, *In a Different Voice: Psychological Theory and Women's Development* (Cambridge, Mass.: Har-

BEST COPY AVAILABLE

vard University Press, 1982), and C. Gilligan and G. Wiggins, "The Origins of Morality in Early Childhood Relationships," in J. Kagan and S. Lamb, eds., *The Emergence of Morality in Young Children* (Chicago: University of Chicago Press, 1987).

28. Alverno College Faculty, *Assessment at Alverno College*, rev. ed. (Milwaukee, Wis.: Alverno College, 1985), p. 1. (Other material is also available from the college.) For a comprehensive analysis of the assessment system at Alverno (and such competency-based programs in higher education more generally), see G. Grant and Associates, *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education* (San Francisco: Jossey-Bass, 1979).
29. Ibid., p. 1.
30. Office of Strategic Services, *Assessment of Men: Selection of Personnel for the Office of Strategic Services* (Troy, Mo.: Holt, Rinehart & Winston, 1948), p. 28 (emphasis added).
31. R. Edgerton, "An Assessment of Assessment," in Educational Testing Service, eds., *Assessing the Outcomes of Higher Education*, Proceedings of the 1986 ETS Invitational Conference (Princeton, N. J.: Educational Testing Service, 1986), pp. 93-110.
32. Office of Strategic Services, *Assessment of Men*, p. 53.
33. American Psychological Association, "Standards for Educational and Psychological Testing" (Washington, D.C.: American Psychological Association, 1985), p. 54.
34. W. Haney and G. Madaus, "The Evolution of Ethical and Technical Standards for Testing," in R. K. Hambleton and J. N. Zaal (eds.), *Advances in Educational and Psychological Testing: Theory and Applications* (Norwell, Mass.: Kluwer, 1991).
35. L. J. Cronbach, *Essentials of Psychological Testing*, 5th ed. (New York: HarperCollins, 1990), and R. A. Berk, ed., *Performance Assessment Methods and Applications* (Baltimore, Md.: Johns Hopkins University Press, 1992).
36. See D. Riesman's insightful history, "Encountering Difficulties in Trying to Raise Academic Standards," in Grant and Associates, *On Competence*.
37. See, for example, the various papers from a recent conference

on performance assessment in the professions: Educational Testing Service, ed., *What We Can Learn from Performance Assessment for the Professions*, Proceedings of the 1992 ETS Invitational Conference. (Princeton, N.J.: Educational Testing Service, 1993).

38. See, for example, Elbow, "One-to-One Faculty Development"; Elbow, "Trying to Teach While Thinking About the End"; and R. Nickse and others, *Competency-Based Education* (New York: Teachers College Press, 1981).
39. D. McClelland, "Testing for Competence Rather than for 'Intelligence,'" *American Psychologist* 28 (1973): 1-14.
40. See, for example, F. Smith, *Insult to Intelligence: The Bureaucratic Invasion of Our Classrooms* (Portsmouth, N.H.: Heinemann Educational Books, 1986).
41. See G. Wiggins, "Standards, Not Standardization: Evoking Quality Student Work," *Educational Leadership* 48 (1991): 18-25; compare G. J. Cizek, "Confusion Effusion: A Rejoinder to Wiggins," *Phi Delta Kappan* 73 (1991): 150-153.
42. Frederiksen, "The Real Test Bias," p. 201.
43. Cronbach, *Essentials of Psychological Testing*, pp. 74-75.
44. Haney and Madaus, *The Evolution of Ethical and Technical Standards for Testing*, p. 34.
45. M. Feuer, K. Fulton, and P. Morrison, "Better Tests and Testing Practices: Options for Policy Makers," *Phi Delta Kappan* 74 (1993): 530-533, p. 532.
46. Ibid.
47. Department of Education, *Assessment for Better Learning: A Public Discussion Document* (Wellington, New Zealand: Department of Education, 1989).

2 Assessment Worthy of the Liberal Arts

Assessment of *what?*[1] Assessors cannot exercise judgment or use any tools unless they are clear about the criteria and standards to be used in the judging. In other words, we must determine the aim of education before we can determine what we should assess and what kind of evidence we should seek.

Despite what many testing programs implicitly assume, the aim of education is thoughtful action, not "knowledge." Our syllabi are means to a larger end: developing the "disciplines" that are at the heart of each discipline. A capacity for autonomous learning and a thirst for unending education are more important than accurate recall or simplistic application of the particular knowledge taught. The implications for assessment are fundamental: we need to be assessing primarily for mature habits of mind and a thoughtful and effective use of knowledge.

Consider, therefore, the first known assessor of intellectual achievement. I am thinking, of course, of Socrates—the Socrates of the dialogues of Plato, in which we regularly see those who either appear to be or profess to be competent put to the "test" of question, answer, and (especially) sustained and engaged conversation. Socrates the assessor: he is certainly an odd one by conventional standards. He does not seem to have nice answer keys or scoring rubrics by his side. And his dialogues never lead to "knowledge" or the kind of closure favored by traditional teachers and assessors. Rather, what is at stake is determining whether those with whom Socrates

speaks have the right habits of mind—the "disciplines" of the liberal arts.

These aims and methods can be seen through the Platonic dialogue called "Meno."[2] Meno, a brash young fellow, comes up to Socrates and abruptly asks a question. The first lines of the dialogue are: "Can you tell me, Socrates, whether virtue can be taught, or is acquired by practice, not teaching? Or if neither by practice nor by learning, whether it comes to mankind by nature or in some other way?" Meno apparently needs to know—now. Socrates responds in a very annoying and typically Socratic way. He says that he cannot answer the question because he does not know what virtue is. Meno is clearly astonished to learn that a bona fide, certified sage does not know what *everybody* knows—namely, what it means to be good. But after Meno makes the foolish mistake of venturing to tell Socrates what virtue is, Socrates proceeds to undress him two or three times.

Finally, in exasperation at having his own accounts of virtue turned inside out and found wanting, Meno says something that goes to the heart of the distinction between conventional testing and an assessment worthy of the liberal arts. Meno says, "Well now, my dear Socrates, you are just what I have always heard before I met you. Always puzzled yourself and puzzling everyone else. And you seem to me to be a regular wizard. You bewitch me. You drown me in puzzles. Really and truly, my soul is numb. My mouth is numb. And what to answer you I do not know." For our concerns, it is the next line that is most important: "Yet I have a thousand times made long speeches about virtue before many a large audience. And good speeches too, as I thought. But I have not a word to say at all as to what it is."

Meno's ironic comment highlights the difference between merely dutiful learning in school and real intellectual excellence. Meno is reduced to speechlessness, he thinks, because of the sophistry of Socrates' questions and analyses; the thoughtful reader knows, however, that Meno does not really know what he is talking about. He is unable to move beyond clichés, he fails to see the contradictions in his diverse arguments, and he cannot justify his opinions when asked. Yet the dialogue presents Meno as a conventionally successful student. How do we know? The real-life Meno

BEST COPY AVAILABLE

was, in fact, a successful young military man. And throughout the dialogue, Meno is constantly dropping references—the ancient equivalent of student footnotes—to all the famous people who say this and that about virtue (and whom he, of course, agrees with). It may be true, as Meno claims, that he can be a successful speaker—passionate, convincing.

One of Plato's intentions here, then, is to challenge the views of the Sophists: competent presentation is not adequate evidence of intellectual mastery; rhetorical skill, using borrowed ideas, is not understanding. Mere learnedness or eloquence is not what a liberal education is about. As Plato has Socrates say elsewhere, education is not "putting sight into blind eyes" but liberating the mind from mere opinion and hackneyed thinking.

Meno is really like so many students—a memorizer, able to make effective speeches with references to famous people, sayings, and works. We are meant to know that his name is used throughout the work as a pun. It is very close in Greek to the word for memory: μενον (Meno), μενων (power). Is that not what too much of our assessment is already about? Do we not too often fail to assess whether the student can do anything more than cite borrowed quotes, arguments, facts, and figures?

We also know from history that the real Meno was a nasty fellow: clever, effective, ruthless. It was no coincidence, then, that Plato titled a dialogue about morality and education as he did. He clearly meant for us to recall Meno's actual character while hearing his mind at work. Meno's impetuosity in the conversation and his desire to merely buttress his unexamined opinions with what wise men say make a dangerous combination. Plato wants us to see, through Socrates, that conventional education can be quite dangerous. As we get better and better at our lessons and our craft, we may become less and less likely to question what we know.

We are meant to see that there is an inescapable moral dimension to all learning (even abstract academic learning) and to our assessment of its success. An education is not merely a training; skill can be used for good or for ill, thoughtfully or thoughtlessly. A thoughtful assessment system does not seek correct answers only, therefore. It seeks evidence of worthy habits of mind; it seeks to expose and root out thoughtlessness—moral as well as intellectual

thoughtlessness. Sometimes, therefore, it is not factual error but the student's *response* to error that reveals either understanding or lack of it.[4] Focusing squarely on such habits of mind as openness to ideas, persistence, and willingness to admit ignorance makes different demands upon the student than focusing on the test criteria of the current system—and different demands upon the assessor as well. At the very least, judgment must be seen to be an essential element of assessment: mere right answers to uniform questions are incapable of revealing its presence or absence.

"Thoughtless mastery" (as I have elsewhere termed it) really does exist; it is not a contradiction in terms, though typical testing cannot easily discern thoughtful from thoughtless mastery.[5] For what must be assessed is not whether the student is learned or ignorant but whether he or she is thoughtful or thoughtless about what has been learned. Our assessments tend unwittingly to reinforce thoughtless mastery as an aim by failing to distinguish between "thoughtful use" and "correct answer" and by routinely valuing de facto the latter. A now-famous test question from the National Assessment of Educational Progress (NAEP) in mathematics, discussed by Alan Shoenfeld of Berkeley, makes this clear: "'An army bus holds 36 soldiers. If 1128 soldiers are being sent by bus to their training site, how many buses are needed?' Of the students who worked the problem 29% wrote that the number needed is '31 remainder 12' while only 23% gave the correct answer."[6]

Unthinking habits run deep: a steady dose of decontextualized problems with right numerical answers leads students to not question the idea of a "remainder 12" bus. We dare not too quickly blame the student-victim. This is not mere carelessness on the student's part. It is learned thoughtlessness, induced by an assessment system composed of tests containing decontextualized items whose answers have no real consequence or real-world meaning and that are not meant to be lingered over.

Eight Dilemmas of Student Assessment

This tension between unthinking learning and thoughtful (re)consideration is the first of eight dilemmas at the heart of the assessment of intellectual progress. Much learning paradoxically *requires*

BEST COPY AVAILABLE

unthinking learning of pat formulas, drills, and declarative statements. But understanding is something different than technical prowess; understanding emerges when we are required to reflect upon achievement, to verify or critique—thus to rethink and relearn—what we know, through many "tests" of experience, action, and discussion (knowledge-in-use). Understanding involves questioning, as Socrates so often did, the assumptions upon which prior learning is based.

Consider physical mastery. One of my passions is baseball, and in George Will's wonderful book *Men At Work*, on the craft of playing and managing major-league baseball, there is an odd but insightful phrase, "muscle memory," that well describes thoughtless prowess. Good hitters talk about not thinking too much at bat. What has to take over the hitter is "muscle memory"—a wonderful phrase for the kind of unthinking skill that we admire. On the other hand, if hitters whose batting average is falling want to genuinely understand hitting and alter a set of habits that no longer serve them, they must study hitting, analyze it, and reconstitute it in the form of new habits—"second nature," in that apt phrase—using a feedback loop of constant assessment and adjustment to ensure that the new habits work. Similarly, I do not want my brain surgeon to be thinking about what health really is while I am under the knife. But to truly understand and honor the Hippocratic Oath, every doctor must repeatedly undertake deep meditation on the nature of care. And to avoid unthinkingly seeing my case as merely a subset of a predictable condition, the doctor needs to be responsive—possessing what Freud called "free-floating attention."

We can put this first dilemma in the language of relationship. Artistry requires habit-induced skill but also tact and good judgment, the nontechnical inclination to be alert to the perhaps unique elements of the present situation. No assessment system worthy of the liberal arts assesses mere knowledge or skill; it assesses intellectual character, good judgment—the *wise* use of know-how in context. We are therefore derelict if we construe assessment as only the act of finding out whether students learned what we taught. On the contrary, one could argue that assessment ought to determine whether students have effectively reconsidered their recently learned abstract knowledge in light of their experience. (This

is why the case method and problem-based-learning methods in law, business, and medicine are now so common and apt.)

Schooling aims to transmit what is known; we should thus assess for knowledge of what was taught. But schooling also aims at intellectual autonomy and the generation of future knowledge and personal meaning; we must assess, then, for what may be generative of future knowledge (which may have little to do directly with testing what was taught). Therein lies our second dilemma in assessment: we must worry about what students know, but we must also worry about whether what they know now has any meaning. To put this in old, familiar language: the higher-order act of synthesis in the Taxonomy is a creative act of connection making; synthesis is *not* necessarily achieved or better assessed after knowledge is exhaustively taught and assessed (that is, by marching through the Taxonomy in chronological fashion, something against which Bloom in fact vainly argued.)

We certainly say we would like to see more "real" thinkers, and we bemoan doltish behavior in our students, but I think we do protest too much. Our testing and grading habits, which give us away, show that we do not negotiate the dilemma well. Look how often, for example, students give us back precisely what we said or they read. It is too easy to come to school and college and leave both one's own and the teacher's prejudices unexamined. If in so doing students "succeed" (get A's), they are left with the impression that assessment is merely another form of jumping through hoops or of licensure in a technical trade. And yet school and college offer us the socially sanctioned opportunity, indeed the *obligation*, to disturb and challenge students intellectually, using knowledge as a means, not as an end, to engender more effective and self-reflective thought and action. That view is at odds with the habit of using uniform tests of the content of the syllabus.

Do not make the mistake of hearing this last point as an ode to evaluation-free schooling. On the contrary, I think effective education is impossible without constant, rigorous assessment that results in good feedback. But we should require more than rigorous testing, as I said in Chapter One. We should require respectful and responsive forms of assessment, in accordance with our deeper objectives.

We have the obligation to "test" and "examine" each student's supposed achievements. But we must do so in a way that is ennobling, fair, and responsive to the student's intellectual needs. This is our third dilemma, introduced in the previous chapter: a test is designed to standardize and simplify our assessment, but to rely on tests is to ensure that students' "knowledge," not their intellectual progress as thinkers and producers, gets assessed. A more enlightened concern for students' need to "show off" what they know (to use Theodore Sizer's fine phrase) would make us more sensitive to their need for regular opportunities to *shape the terms of evidence of mastery*—a contract between the student and the assessor to deliver evidence of progress that is mutually acceptable. This is a dilemma because it wreaks havoc with efficient testing of many students at once and with the possibility of comparable assessment results—values that we also hold dear.

The fourth dilemma involves the tension between our need to probe and our obligation to be respectful. To "examine" what the student knows requires me to be mistrustful, in a sense. I put the student's knowledge to the test; I suspend my naive faith in the student's apparent mastery. What does the student *really* know? My aim is to ferret out not only what the student knows but what the student seems or claims to know but in fact does not—as Socrates did with Meno and all the other discussants. Compounding the ill effects of our inherent mistrust, we often employ many disrespectful practices: we employ secrecy about what is tested and graded, for example, we use tricky questions and answers ("distracters"), and we rarely give the student adequate opportunity to justify an odd or seemingly incorrect answer. (These practices are discussed more fully in Chapter Four.)

We must replace these practices with the tact of Socrates: tact to respect the student's ideas enough to enter them fully—even more fully than the thinker sometimes—and thus the tact to accept apt but unanticipatable or unique responses. Like Socrates, we must be prepared to give up control of the conversation and the order of our questions. That is what a dialogue is, after all: a mutually respectful and responsive (hence nonscriptable) shared inquiry. So often we are disrespectful in the sense of having one question, one sort of

answer, and one protocol in mind. We rarely treat the student as a respected colleague.

Whatever the lure of standardization, therefore, it is opposed to the flexible settings that evoke personalized understanding. Establishing high standards that are adapted to each child and circumstance of assessment is very difficult: How many of us can be a Socrates, a Piaget, a Freud? But at what cost to *our* understanding—hence to validity—do we settle for held-in-common answers—those that are the most easy to obtain?

That the tension between rigor and respect, standardization and humanely held standards is real can be seen in the various uneasy compromises played out in traditional schools and alternative schools. "Good" schools have teachers who "grade hard" (and on a steep curve) and test a lot. But with what rationale and at what cost? In one "good" district in which I worked, science teachers gave to their seventh- and eighth-graders a final exam that contributed 15 percent of the final grade. The questions—over 100—were picayune, and the test was unreliable: the average test grade was a full letter grade lower than the year's grade for students of all abilities. What does the student take as a lesson from such a system about academic rigor, academic values, and fairness?

On the other hand, I often hear the wish of faculties to do away with formal evaluation, tests, or grades—sometimes even in these "good" schools—in the supposed interest of students. Some alternative-school faculties seem to go further, viewing all formal assessment and grades as antithetical to their mission—as if competence could somehow be developed without formal feedback. These well-meaning folks often end up confusing effort with achievement, individual growth with genuine progress.[6] If I had to choose between, on the one hand, mickey mouse "gotcha!" tests with norm-referenced scoring and grading and, on the other hand, an absence of uniform tests and grades, I *might* go with the alternative schools; but it is a bad choice, and it shows that we have not understood and negotiated the dilemma.

So we must think more carefully about how to balance the nurturing of diverse intellectual urges with the need for maintaining and inculcating standards—a quest for humane yet effective rigor, standards without mere standardization. In the proper drive

for more authentic forms of assessment, we have to ensure, for example, that the cry does not lead to mere "projects" with no clear or valid standards. We must do something more than simply find a more hands-on way of engaging students; these new performance tasks must provide clear insights and discriminations about who has and has not mastered essential tasks and outcomes.

A fifth dilemma is that, despite our desire to test what was learned in a dispassionate fashion, no test is ever neutral—about epistemology. Tests *teach*. Their form and their content teach the student what kinds of challenges adults (seem to) value. If we keep testing for what is easy and uncontroversial, as we now do, we will mislead the student as to what work is of most value. This is the shortcoming of almost all state testing programs. Generations of students and teachers alike have fixated on the kinds of questions asked on the tests—to the detriment of more complex and performance-based objectives that the state cannot possibly test en masse. A person from India described a more harrowing result: when teachers try to do interesting things in their secondary classes in India, he has heard students protest, in loud voices, "Not on the test! Not on the test!" In other words, students then easily come to believe that knowledge is a repository of sanctioned information instead of a "discipline"—a flexible and context-sensitive way of thinking and acting.

Furthermore, tests *should* teach the student, not only about how school tests parallel the "tests" of professional knowledge and understanding but about the questions and challenges that define a given field. Otherwise, an unrelenting dose of efficient, "proxy" forms of assessment teach the student misleading lessons about what an intellectual test is, and they reduce the likelihood of marginal students being interested in intellectual work.

Testing that teaches what we *ought* to value is technically difficult, time-consuming, and dependent upon the kinds of sophisticated task analysis that teachers have little time for. Therein lies the dilemma: few teachers and school systems can imagine what a comprehensive "examining" system might look like—a system that gives each student many options for assessment and that "tests" the most complex aspects of performance and production. Such approaches are more routine in other countries. The Italians, for example, require all students to take oral exams, even in mathematics. Victoria, in Australia, makes locally generated student work part of the state assessment. The International Baccalaureate includes an "extended essay" as a graduation requirement; the work is judged by disinterested readers. Some U.S. schools, particularly the more progressive or alternative, have similar requirements—most notably, Central Park East Secondary School in New York and Walden III in Racine, Wisconsin, both of which require elaborate portfolios and orals for graduation. All of these entities have had to make sacrifices and readjustments to run their assessment systems. Ultimately, of course, what we assess is what we *really* value.

A sixth dilemma inherent in assessing student achievement is balancing testing for the mastery of the ideas and products of other people against testing for the mastery of one's emerging ideas and products. Why is this a dilemma? Why not stress both? Elementary and graduate teachers often do, as I noted in Chapter One. But because an education that centers on doing one's *own* work well is inherently idiosyncratic, the instruction and the assessment cannot, in principle, be standardized in the conventional sense. An unfortunate teacher habit also enters here: too many people believe incorrectly that students must gain control of common knowledge *before* they can be creative with or critical of knowledge. To paraphrase Thomas Kuhn's view, one must have complete control over the existing "paradigm" if dramatic new paradigms or original thoughts are to occur.[7]

Whatever Kuhn's merits as a historian and philosopher of science, I think he is dead wrong about education. I think it is vital to ensure that students immerse themselves, from the word go, in pursuing their own work and questioning established "truths" as they learn. Otherwise, they may have a *long* wait, and their critical judgment and imagination may atrophy. Many bright and able minds drop out mentally or physically because they cannot wait so long for intellectually stimulating challenges of that sort. And the ones that *do* stick around may be more dutiful than thoughtful.

Inevitably, if we first demand knowledge of the orthodox facts and opinions, we run a moral as well as an intellectual risk: the risk of letting students believe that Authority and authoritative answers matter more than inquiry. We may well end up convincing

students that "knowledge" is something other than the result of personal enquiries built upon questions such as theirs. And many students *do* believe that there is "knowledge" over here and there are "questions and ideas" over there, and never the twain shall meet.

The problem can be reframed as a classic one and put forth as our seventh dilemma: the distinction between utilitarian and nonutilitarian (or liberal) knowledge. We are in fact not training historians, scientists, or mathematicians. Too many teachers act as if they were providing technical training to apprentices. There is an important sense in which the liberal arts *are* useless, summed up in that comment supposedly made by Euclid 2,000 years ago when someone complained that geometry was not good for very much. He said, well, give him three drachmas if he has to get him some usefulness out of the study. Schooling is not the same as trade school.

But there is a more important truth in students' desire for an education with more apparent usefulness. We often fail to hear this lament for what it is: a request for more *meaning* in their work (as opposed to relevance). There are many instructional and assessment challenges that are not relevant to students' immediate practical concerns and interests. They nonetheless offer enticing, insightful, and ennobling challenges. For example, I have watched a class of geometry students happily spend two hours on the question, Is the Pythagorean theorem true if one uses shapes other than squares on the legs of the triangle? Adults often disappoint in the other direction, by pandering to students—that is, pursuing ideas that are *too* relevant because they are transitory and ultimately without either meaning or significance.[8] Test and syllabus designers are perpetually insensitive to students' need for genuine problems to chew on, not packages of predigested "knowledge" or artificially simple and unengaging drills.

The dilemma about the liberal arts can be restated in a way that makes clear why a liberal education is problematic. We all need to feel competent, and we want to believe that our education is providing us with something of value. From school competence comes confidence and, we think, greater clarity and direction about one's future, from such confidence comes greater risk taking and thus even greater competence. The trouble with a really *good* education, however, is that it often fails to satisfy this need for self-

satisfaction and vocational direction in the short term (if at all). We have to recognize that an education that aims at deep understanding may cause anxiety or impatience in students; the urge to shun or resist schooling may be well founded. Our dropouts may actually be the tip of the iceberg, more extreme cases of the psychic dropouts who inhabit all our schools and colleges. (Socrates notes that the man liberated from his chains in the Cave *resists* the steep, rough ascent to intellectual freedom.)

The unending questions at the heart of a good education are always disturbing. Meno, you will recall, warns Socrates not to travel to foreign lands, where people would very likely not take kindly to his strange questions. (The allusion to the fate of the real Socrates is clear.) Most students do not deal well with the ambiguity and uncertainty that are the hallmark of a genuine education: I recall a Harvard medical student, shown as part of a group of students experiencing problem-based learning on an episode of "Nova" on PBS, who said defensively, "I didn't come all this way to spend this kind of money to teach *myself*." It is thus naive and misguided to argue that intellectual standards are self-evidently desirable. Our *best* learners at lower levels (where learning is more rote) may balk at the upper levels. Setting high intellectual standards without providing incentives to persist, opportunities to fail and improve, and the tact, wisdom, and guiding model of a mentor is a cruel and hypocritical stance.

The eighth and final dilemma follows from this tension between setting standards and providing incentives to meet them. We aim to maximize everyone's achievement, but we cannot imagine how to do so without lowering or compromising our standards. We demand excellence from all, but we do not expect it. We establish mission statements under the assumption that all children can learn what we have to teach, but we test and grade under the assumption that some will do well and most will not.

The prejudices that underlie these expectations go very deep—far deeper than any contemporary biases, as Michel Foucault reminds us in describing the constant, elaborate, and deliberate ranking of pupils within French schools in the seventeenth and eighteenth centuries.[9] We may set out initially to uphold high standards, but many of our assessment practices unwittingly exag-

gerate differences and become self-fulfilling prophecies about our inability to do so. The standard curve could have been used as often as it is only under the assumption that uniform excellence due to deliberate education is impossible. All one need do to see the error of such thinking is to note the across-the-board quality performance that occurs through military, athletic, musical, or dramatic training in the best programs or to remember that in the original mastery learning research, the key finding was that "equal time spent on task" is a more important variable than aptitude for yielding a wide range of performance results; varying the learning time allowed leads to a significant decrease in range of results and often to across-the-board mastery of material.

On the other hand, the evidence that uniformly high standards can be upheld and met by everyone is usually dependent upon external assessments of some kind, as the examples suggest: we are delighted when everyone gets a 5 on an Advanced Placement exam. Yet educators *within* "good" schools are convinced that standardization in assessment will lower standards, and the minimum-competency movement seems to bear out their fears. On the other hand, idiosyncratic (and often eccentric or perverse) testing and grading easily result in schools and colleges where external standardization is resisted. This undermines a common concern for quality and student respect for standards. Is it not possible for a faculty to set high standards for all students and also make it possible for all students to meet those standards without charges of fraud or perversion of standards? (Is that not what the best colleges and private schools do?) Is there, in fact, such a thing as *one* standard? Could one not argue that there are as many standards as there are aspirations and programs?

Some Postulates for a More Thoughtful Assessment System

Our task, then, is more difficult than we perhaps first imagined. The quest is not for "better" tests, as if a mere technical deficiency or ignorance lay at the heart of our current problems in assessing for ultimate educational aims. We have to bring to consciousness and thoughtfully examine our deep-seated habit of seeking superficially correct answers to uniform questions. Technical improvements in performance testing will never obviate the need for careful judgment in the use of such testing, because the dilemmas just discussed are unavoidable. Are testers not therefore obligated to more carefully consider the dangers and limits of their techniques? How might we more effectively grasp and negotiate these dilemmas? Let me offer nine postulates, with some examples for each, as a way of considering these dilemmas and questions more carefully.

BEST COPY AVAILABLE

Postulate 1: Assessment of thoughtful mastery should ask students to justify their understanding and craft, not merely to recite orthodox views or mindlessly employ techniques in a vacuum.

We suffer from an impoverished conception of what it means to *know* something. Understanding is not displayed by "correct" answers to questions and problems out of context; on the contrary, misunderstanding is easily hidden behind thoughtless recall. Mastery of the liberal arts is not a mastery of orthodoxy but an ability to effectively *justify* one's answers and arguments—to a real audience or client, and in specific contexts where generic answers are not adequate.

Our schools, but especially our universities, are schizophrenic in this regard. Their traditions often reveal their roots in the rigid religious training of premodern times. But they now exist to foster research. There was and is an irresolvable tension between promoting orthodoxy and promoting inquiry. Whatever our modern ideology about inquiry, we still lean pretty heavily on the orthodoxy side in our assessment: up until the graduate experience, students have to first demonstrate their control over other people's knowledge in all subject matters. And K-12 schooling is filled with assessments that require the student to master an orthodox format (for example, the five-paragraph essay or three-step math proof), irrespective of the quality of the student's thinking. But this academic orthodoxy has little to do with the ultimate aim of developing effective inquirers, researchers, and scholars as exemplified in the ultimate educational "test"—the dissertation and oral in defense of a thesis.

A dissertation, high school history term paper, or third-grade book report does not provide adequate evidence of mastery; it is the *defense* or *discussion of* that product that reveals whether or not

understanding is present. We must remember that all assessment should point toward the last stages of education and our assessments in the final stage (original research and defense against critics), so as to give the student both frequent warning of the need to *justify* opinions that are developed as a result of teaching and frequent opportunities to do so.

An effective examination is most certainly not "gotcha!" testing. The assessor is meant to probe, reframe questions, and even cue or prompt an answer, if necessary, to be sure of the student's actual ability and to enable even a poor performer to learn from assessment. In many orals, for example, the first answer (or lack of one) is not deemed a sufficient insight into the student's knowledge.[10] Not only is a first answer not decisive in the evaluation; there is a clear commitment to determine what it is the student really understands underneath apparently wrong or odd answers. The student is properly *rewarded* for self-corrections and self-conscious clarification of earlier responses.

Speaking logistically, a wholesale move toward more oral examinations would be difficult (though they are a routine part of subject exams in other countries). But there are other, less time-intensive measures that are feasible—measures that would allow us to honor the obligation to examine students' responses to follow-up questions and probes of their ideas, not merely to note and evaluate their first answers. The obligation implies, for instance, that in assigning a paper and evaluating it, the student should have to respond to our criticism (or the criticism of some other audience, perhaps of peers), to which we then respond *as part of the formal assessment process*—not as a voluntary exercise after the test or paper is completed.

To teach students that we are serious about intellectual standards, we must always assess their ability to see the limits of what is learned; they need to have the chance to punch holes in our own or the textbook's presentation. They have a right to demand justification of our point of view. That is what a liberal education is about. It also sends the right moral message: we are both, student and teacher, subservient to rational principles of evidence and argument.

Postulate 2: The student is an apprentice liberal artist and should be treated accordingly, through access to models and feedback in learning and assessment.

Any novice or apprentice needs models of excellence, the opportunity to imitate those models, and the chance to work on authentic scholarship projects.[11] Students should be required to recognize, learn from, and then produce quality work in unending cycles of model-practice-feedback-refinement. They should not get out of our clutches until they have produced some genuinely high-quality work of their own.

As I mentioned earlier, the International Baccalaureate (I.B.) has such a requirement—an "extended essay" involving student research in any I.B. subject. Students explore such diverse topics as the physics of a golf ball or the images in Jamaican poetry. (The I.B. even publishes a high-quality book containing some of the best essays from around the world.) What is instructive and gratifying about the I.B. assignment is the insistence that the student have a personal stake in the research. The guidelines explicitly discourage the kinds of sweeping, overambitious (and hence superficial and uninteresting) papers that our students too often produce. For example, under the guidelines for economics and for literature, the following advice is offered: "An unacceptable essay is one in which there is no personal research, which is dependent entirely on summarizing secondary sources. . . . Encourage: 'Price Theory and hairdressing in my town.' Discourage: 'OPEC 1980–1990.'" "Candidates should avoid topics which are general or vague. These usually lead to a superficial survey, often borrowed directly from textbooks, which has little or no educational value. Examples of topics to avoid: the origins of Romanticism, the use of nature in poetry, Greek mythology in English literature, etc."[12]

An apprentice must see and understand progress in a given craft—in this case, knowledge production. Paradoxically, that means, in part, learning to see one's past standards as now unacceptable. One of my favorite assignments when I taught at Brown was to ask students in their final paper for a rewrite of their first paper, based on all that they had since learned or thought. A number of the seniors told me that it was the most important event in their four years. They were astonished to see how their thinking

had changed and to discover how sloppy their "complete" work seemed to them in retrospect.

Further, they were learning that thinking does not stand still—and that it *should* not. In demanding intellectual excellence of novices, we begin to focus our assessment on what Aristotle called the "intellectual virtues." Does the student display a sense of craftsmanship, perseverance, tolerance of ambiguity? Can the student display empathy when everyone else is critical, or be critical when everyone else is empathetic? Can the student, *without prodding*, rethink and revise a paper or point of view? An education is ultimately about those intellectual virtues. When all of the knowledge has faded away, when all of the cramming has been forgotten, if those intellectual dispositions do not remain, we have failed.

While some people get very squeamish about assessing such things as perseverance, style, craftsmanship, and love of precision, I do not. If we value something, we should assess it. The best way to assess such habits is indirectly: to devise tasks that require them, tasks that can be done well only if the requisite habits are present and well tapped by the student.[3] As I noted above, this indirect assessment was used by the OSS in testing spy candidates. All assessment would be much improved, in fact, if we thought of it as a kind of intellectual Outward Bound. It should never be possible to do an end run around those desirable habits. Students who can get A's by missing class, cramming, or articulateness and native ability only are telling us something about the failures of our assessment system.

Sometimes improved assessment involves as subtle a shift as sending the message day in and day out that quality matters—not just because teachers say so but because the situation demands it. Consider one simple strategy purportedly used by Uri Treisman at Berkeley in his successful work with minority mathematics students. He demands that every piece of work that students hand in be initialed by another student; students get both the grade for their own paper and the grade for the paper on which they signed off. This grading policy makes it clear that one is responsible for one's work, as both a producer and an editor; there are consequences for failing to adequately self-assess one's work or critique the work of others properly. One can go further by designing tests in which

situational consequences occur through one's success or failure: if the problem in physics involves a model bridge needing to withstand a certain load; if the persuasive essay genuinely has to persuade a professional editor; if the student studying German has to order specific merchandise and request information from a German firm, then grades become apt symbols for real qualities and consequences. (This kind of "authentic simulation" and "quality control" is possible only when we also provide students with useful feedback as part of the assessment process, as we shall see in Chapter Six. *All* assessment should be thought of as "formative," to put it glibly.)

Postulate 3: An authentic assessment system has to be based on known, clear, public, nonarbitrary standards and criteria.

The student cannot be an effective apprentice liberal artist without models. There is no way to empower the student to master complex tasks if the tasks, criteria, and standards are mysterious (or are revealed, but not in advance). It is no wonder, then, that the ubiquitous one-shot, secure test and the often-secret scoring criteria undermine our educational aims.

When we look at the performance world (as opposed to the academic world), we see how much easier it is for performers to be successful, because the "tests" are known from day one. The sheet music, the script, the rules of debate, the rules and strategies of the game are or become known: genuine mastery involves internalizing public criteria and standards. Unfortunately, in education, especially in higher education, the primary vestige of our medieval past is the use of secret tests. (The novices always had to *divine* things.) I was disappointed to learn, when I was a teaching assistant at Harvard, that undergraduates are still not officially allowed to see their blue books after the final exam. But it could be worse: I was told that at Oxford and at Cambridge, they burn blue books!

This unfortunate and deadly tradition is a legacy of tests used as mere gatekeepers or as punishment/reward systems, not as empowering and educative experiences designed for displaying all that a student knows. Most people would likely say, if asked, that it is the *student's* responsibility to figure out what will be tested, not the teacher's responsibility to make it unambiguous. But why would we

not require the school or university to meet students halfway and give them a chance to play from their strengths?

Possible solutions include strategies as simple as giving students the option of alternative forms of the same assignment or handing out in advance a long list of possible questions from which a few will be chosen for the exam (a practice common in colleges and graduate schools). Other solutions include the supplying in advance of scoring rubrics, model papers, or videotaped model performances—anything that would give students an insight into the standards in force. Here is an example of a scoring rubric from a past Advance Placement (AP) exam in U.S. history:

> *Question:* "The economic policies of the federal government from 1921 to 1929 were responsible for the nation's depression of the 1930's." Assess the validity of this generalization.
>
> *Scores*
>
> 13-15 An accurate, well-written response that directly assesses the validity of the generalization. Demonstrates a clear understanding of governmental economic policies; for example, tariffs, pro-business legislation, and foreign debt. Uses at least three specific examples or covers many topics with an intelligent conclusion.
>
> 10-12 A good answer that attempts with some detail to assess the validity of the statement, if only implicitly. Should cover at least two areas of economic policy, but may contain a few minor errors of fact.
>
> 07-09 A reasonably coherent discussion, but with little analysis of economic issues. Answer is not fully developed; may discuss only one issue beyond the level of assertion; for example, laissez-faire policies. Or may give a coherent discussion of concepts without citing specific acts or policies.

> 04-06 Little if any assessment of the statement. An overgeneralized answer, without supporting evidence. . . . The stock market crash must be seen as not merely an event but a consequence of prior policies. . . .[14]

It is expected that the AP teacher will have gone over past essay questions and scoring systems with the student (indeed, books are available from the AP program in which past questions, scoring rubrics, and sample essays are provided, with commentary). Note, however, that the language of the rubric is *inherently* vague, representing as it does the generalizations of strengths and deficiencies found in the many different papers judged by readers to be of equivalent value. The student, to understand and effectively use the rubric, needs to see samples of the papers that correspond to the scores.

What is less noble in the AP experience, of course, is the fact that the student has neither advance knowledge of the question that will be asked nor access to resources in answering the question. What, then, are we really assessing here, insofar as the student cannot be expected to be an expert on every conceivable question that the AP examiners might propose? I pursue this problem thoroughly in Chapter Seven, where I argue that "authenticity" often depends more on the constraints of the test and its administration than on the task itself.

Postulate 4: An authentic education makes self-assessment central.

The means for dispelling secrecy are the same means for ensuring a higher across-the-board quality of work from even our most worrisome students: teaching students how to self-assess and self-adjust, based on the performance standards and criteria to be used. Alverno College, mentioned earlier, has successfully made self-assessment central to its program. In one of my favorite examples, assessment of the communications competency, a student must early on give a videotaped talk. One's first hunch might be that it is the talk that is going to be assessed. No: after the student gives the talk and it is videotaped, the student is assessed on the accuracy

of her self-assessment of that videotaped talk![15] If we want people to gain control of important habits of standards and habits of mind, then they have to know, *perhaps first of all*, how to accurately view those things and to apply criteria to their own work; they cannot always be dependent upon another person for assessment. (Here again we see the importance of not treating Bloom's Taxonomy as a chronology for teaching: "evaluation" is taught from the beginning at Alverno and in other competency-based programs.) Habit development depends upon constant self-assessment, but you need to know what you are *supposed* to be doing before you can do it.

One practical implication of this postulate is that we should require students to submit a self-assessment with all major pieces of work. (Alverno requires each student paper to have clipped to it a self-assessment and one of the subscores given by the teacher is for the accuracy of the self-assessment.) The Advanced Placement art portfolio is a different example of such a system: students submit both a letter and a set of works to the reviewers. The letter explains their intent; the works reveal the actual effect. The judges determine whether, in their professional judgment, the students' intentions were fully realized in the work in question. (This also shows how it is possible to score work rigorously where the work submitted by a group of students is not superficially comparable.)

Postulate 5: We should treat each student as a would-be intellectual performer, not as a would-be learned spectator.

Most courses (and almost all traditional tests) treat the student as a would-be learned spectator rather than as an apprentice intellectual performer. The student must metaphorically "sit in the bleachers" or do drill on the sidelines while others (professors, teachers, and writers of textbooks) "perform." But a liberal education aims at the student's ability to employ knowledge effectively and gracefully, in the context of authentic problems and interactions.

Too many schools define *mastery* as accurately remembering and applying what others say in a plug-in sort of way. This anti-intellectual posture easily induces passivity and cynicism in students. To unendingly postpone the students' doing of their own work is to turn powerful ideas and challenges into drudgery. In an

education aimed at would-be performers, on the other hand, students experience the "tests" that face the expert in the field right from the start—having to find and clarify problems, conduct research, justify their opinion in some public setting—while using (other people's) knowledge in the service of their own opinion.

One of the finest classes that I have ever seen taught at any level, which illustrates this point, was at a high school in Portland, Maine. A veteran teacher offered a Russian history course for which the entire syllabus consisted of a series of chronological biographies. It was then each student's job to put the course together, by becoming each person, in turn, and in two senses: through a ten-minute talk and then through an interactive simulation in character. After four or five students had presented their talks (and been assessed by other students on those talks), they had a Steve Allen "Meeting of the Minds" press conference chaired by the teacher; the "journalists" were the other students. Each member of the panel scored the others on their performance.

A striking thing about the in-character reports I heard in that classroom was their engaging quality. I have sat through my share of dreary student reports. These were as delightful, personalized, and informative as any I have ever heard. In response to my query about why, the teacher said that it was very simple: students knew that there were only two criteria by which they were going to be judged—whether the talk was accurate and (more important) whether it was interesting. How difficult can it be to ask for what we really want? Yet how many rubrics have you seen that put a premium on the interest level of student writing or on the persuasiveness of student presentations? Our assessments are mechanical and disengaged. No wonder student performances often are too.

Postulate 6: An education should develop a student's intellectual style and voice.

As suggested by the previous point, what a liberal artist will be, if he or she has "made it," is somebody who has a style. Somebody whose intellectual "voice" is natural, compelling, and clearly individual. Read the turgid prose that we receive *and accept*, and you can see that we are failing to develop style, voice, and an engaging point of view. (Read our own professional writing in aca-

BEST COPY AVAILABLE

demic journals.) Students are convinced that we want merely the party line and that insights cast in compelling prose are an option, not a requirement.

There are a number of ways to get at this style component of assessment. Some writing assessments now score papers for "voice," not just mechanics and organization.[15] Consider, for example, Exhibit 2.1.—the rubric from a Canadian assessment—noting especially the caution to judges at the bottom.

Another approach might ask students, after they have written a lengthy research paper (with all the requisite footnotes and bibliographical information), to turn the same research into a one-page paper to be delivered, in an engaging and insightful way, to an audience of laypersons such as Rotarians.

Do not misunderstand me. This is not just an aesthetic issue, this issue of style or voice. It is an issue of one's inner voice. It is the serious problem of how to strengthen one's faint intellectual intuition in a sea of loud professorial or textbook opinions, how to nurture the seed of a new idea that is easily crushed if not allowed to grow. This is related to the idea of conscience, and it is no coincidence that Socrates talked about his little voice as his most trustworthy guide.

It is easy, as a student, to lose that little voice. But that voice is not just a "personal" voice, irrelevant to "academic" accomplishment. It is the voice of common sense and of inchoate hunches. It is the voice that can turn around and question the importance of what one has just spent two months working on. In other words, it is the little voice that says, Ah, come on, is this really *that* important? It is the little voice that says, You know, there is probably another way to look at this. It is the little voice that says, I have a feeling that there is something not quite right about what the teacher is saying—what Neil P. Postman and Charles W. Weingartner, in great sixties' fashion, called a "crap detector." It is the little voice that most of us do not hear in our students (or ourselves) unless it is asked for. An assessment should ask for it.

There are ways of assessing such seemingly intangible capacities. I saw an English teacher do essentially what I am describing through a peer editing process. He told his students that they should turn back any paper that was boring or slapdash and mark the exact spot where they began to lose interest. The paper was not "finished"

Exhibit 2.1. Alberta, Canada, High School Leaving Exam: Writing Assessment.

Section I: Personal Response to Literature—Scoring Guide

Thought and Detail

When marking **Thought and Detail**, the marker should consider how effectively

- the assignment is addressed
- the detail supports and/or clarifies the response

5 *Excellent:* An insightful understanding of the reading selection(s) is effectively demonstrated. The student's opinion, whether directly stated or implied, is perceptive and is appropriately supported by specific details. Support is well defined and appropriate.

4 *Proficient:* A well-considered understanding of the reading selection(s) is appropriately demonstrated. The student's opinion, whether directly stated or implied, is thoughtful and is supported by details. Support is well defined and appropriate.

3 *Satisfactory:* A defensible understanding of the reading selection(s) is clearly demonstrated. The student's opinion, whether directly stated or implied, is conventional but is plausibly supported. Support is general but functional.

2 *Limited:* An understanding of the reading selection(s) may be evident but is vaguely demonstrated or is not always defensible or sustained. The student's opinion may be superficial, and support is scant and/or vague, and/or redundant.

1 *Poor:* An implausible conjecture concerning the reading selection(s) is suggested. The student's opinion, if present, is irrelevant or incomprehensible. Support is inappropriate, inadequate, or absent.

INS *Insufficient:* The marker can discern no evidence of an attempt to fulfil the assignment, or the writing is so deficient in length that it is not possible to assess thought and detail.

It is important to recognize that student responses to the Personal Response Assignment will vary from writing that treats personal views and ideas analytically and rather formally to writing that explores ideas experimentally and informally. Consequently, evaluation of the personal response on the diploma examination will be in the context of Louise Rosenblatt's suggestion:

The evaluation of the answer would be in terms of the amount of evidence that the [student] has actually read something and thought about it, not a question of whether necessarily he has thought about it in the way an adult would, or given an adult's "correct" answer. (Rosenblatt, Louise. "The Reader's Contribution in the Literary Experience." An interview with Lionel Wilson in *The English Quarterly 1* (Spring, 1981): 3-12.)

Source: Alberta Education (1993). *1993-94 School Year, English 33 Information Bulletin, Diploma Examinations Program.* Edmonton, Alberta. Reprinted with the permission of Alberta Education.

in the peer review process, as a revised draft, until the peer readers were able to read to the end of the paper without placing that mark. That sort of assessment sends a message to students about writing and its purpose—a message that technical compliance with formal criteria is a means to an end, not our aim as writers.[17]

There is another point to this issue of voice and style. The thing that is so ghastly about academic prose is that one really does sense that it is not meant for any real or particular audience. (Of course, sometimes it isn't, in the sense that rhetorical or aesthetic qualities apparently count for nil, as judged by editors.) It seems to me that if we are serious about empowering students, we must get them to worry about audience in a deeper way. We must demand that their work be *effective*. We must demand that it actually reach the audience and accomplish its intended purpose. There is nothing more foolish, in my view, than saying, "Write a persuasive essay" without making students persuade anybody of anything. So let us set up situations in which the student has to persuade readers, or at least get judged by an audience on more than just accuracy. Even Socrates knew, within the clash of Reason and Rhetoric, that teaching had to be not merely truthful but effective.

Postulate 7: Understanding is best assessed by pursuing students' questions, not merely by noting their answers.

Too often in assessment, we worry about whether students have learned what we taught. This is sensible, of course. But let me take an unorthodox position: such a view of assessment, taken to extremes, is incompatible with the "test" of the liberal arts. One important purpose of those "arts that would make us free" is to enable us to criticize sanctioned ideas, not merely retell what was taught.

A less confrontational way to make the point is to remind ourselves that it is the astute questioner, not the technically correct answerer, who symbolizes the liberal artist. We would do well to recall a point made by the philosopher Hans-Georg Gadamer (with his explicit homage to our friend Socrates), who argued that it is the dominant opinion, not ignorance, that threatens thinking.[18] Ensuring that the student has the capacity to keep questions alive in the face of peer pressure, conventional wisdom, and the habit of our own convictions is what the liberal arts must always be about.

Admittedly, *some* knowledge is required before we can ask good questions and pursue the answers we receive. But if we are honest about this, we will admit that the kind of exhaustive expertise we typically expect of students up front is overkill. After all, children are wonderful and persistent questioners. Indeed, academic types are invariably prone to making the mistake that philosopher Gilbert Ryle called the Cartesian fallacy: assuming that a complete "knowing that" must *always* precede and serve as a condition for "knowing how."[19] No person who creates knowledge or uses knowledge to put bread on the table would ever be guilty of this fallacy. All apprentices and would-be performers learn on the job. Given that, as teachers, we therefore tend to overteach or "front load" knowledge, a good pedagogical rule of thumb is this: teach the minimum necessary to get the students asking questions that will lead to your more subtle goals.

We would do well, then, to think of our task as introducing the student to cycles of question-answer-question and not just question-answer—with the aim of a course being, in part, to make the student, not the teacher or text, the ultimate initiator of the cycle. To continually postpone the students' ability to ask important questions in the name of "mastery" is to jeopardize their intellect. Good judgment and aggressive thinking will atrophy if they are incessantly postponed while professors profess. In any event, the most important "performance" in the liberal arts is to initiate and sustain good question-asking.

Some very practical points about testing can be made out of this esoteric argument. We rarely assess students on their ability to ask good questions. Indeed, we rarely teach them a repertoire of question-asking strategies for investigating essential ideas and issues. If what we assess is what we de facto value (irrespective of what we say), then it should become obvious to students through the demands of the course and our assessment strategies that question-asking is central. Too often, however, our assessments send the message that mastery of the "given" is the exclusive aim and that question-asking is not a masterable skill but a spontaneous urge.

The problem goes deeper. Our scope-and-sequence curriculum-writing (and the tests that follow from it) suggests a very inert and atomistic view of knowing. How, then, will students' naive

understanding of the same subjects across the years of education become developed, tested, and integrated? How will we know whether students are becoming not merely "knowledgeable" but also more sophisticated in their understanding of the same important ideas unless we persist in asking the same important questions over time? Jerome Bruner suggested in his seminal book *The Process of Education* that any subject could be taught in a credible and effective way, at any cognitive level.[20] I would go further: the younger student will never make it to the upper levels of academe without being repeatedly confronted with the most important questions and perspectives on those questions, beginning at the earliest levels of education.[21]

Postulate 8: A vital aim of education is to have students understand the limits and boundaries of ideas, theories, and systems.

To paint the starkest picture of the difference between a "liberal" and a "non-liberal" view of the disciplines, we in the liberal camp might see our task as teaching and assessing the ability to gauge the strengths and weaknesses of every major notion we teach—be it a theorem in math, a hypothesis in science, or a literary theory in English. We need to know if students can see the strengths and weaknesses of theories and paradigms. This would include not only the limits of a theory within a unit or subject but across disciplines, as when we apply the rules of physical science to the human sciences.

A few years back, as part of my work with the Coalition of Essential Schools, I made reference to so-called essential questions as a way of formalizing this idea, and a number of schools (most notably, Central Park East Secondary School) developed an entire framework around the idea.[22] There is no novelty in the concept, however. Bruner talked about "guiding conjectures," and Joseph Schwab, many years ago, wrote about and taught from a similar concept at the University of Chicago.[23] He termed the ability to move back and forth across these questions and limits the art of the "eclectic" and I encourage a return to his essays for numerous suggestions on how to help students explore the merits of sanctioned truths and the boundaries of subject-area insight.

I fear that we no longer know how to teach science (or any

established body of knowledge) as a liberal art. To present the sciences as merely logical and technical is to make it increasingly unlikely that nonscientists will profit from studying science enough to support intelligent science policy as adults (and to make science students insufficiently critical). As it stands now, too little of learning science or other subjects involves doing science and too much to do with mastering orthodox algorithms—learning metaphysics instead of physics, as it were; mastering sanctioned (hence inert) truths instead of learning and being assessed on the truth of the matter: the truths are provisional, yielded by intellectual methods and questions that *transcend* the current results.

I know this weakness in our science students firsthand from my high school teaching days. My best students did not understand, for example, that error is inherent in science and not merely the fault of immature students or poor equipment. (Many believed that when the "big boys and girls" do their measuring, the results are exact.) Nor did many of them realize that words such as *gravity* and *atom* do not correspond to visible "things".

Students can be helped to see the limits of ideas by talking about the history of a subject. We still do a poor job of teaching and assessing students' grasp of the history of important ideas, and yet I know of no method by which inappropriately sacred truths can be more effectively demystified and thoughtfully reconsidered. What questions were Newton and then Einstein trying to answer? What did the first drafts of a history textbook look like, and why were they revised? To ask these questions is to open up a new and exciting world for students. To be smug about our knowledge and to distance ourselves from "crude" and outdated theory is to ensure that we repeat the mistakes of our smug and parochial elders.

Consider the history of geometry, the very idea of which strikes many people as an oxymoron. Many college students are utterly unaware of the problems that forced Euclid to develop an awkward parallel postulate (which was instantly decried by his colleagues). So much for "self-evident truths," that glib phrase found in superficial textbook accounts of Greek mathematics!

The practical consequence of our failure to reveal to students the history of important ideas and assess their understanding of that history is twofold. For one, students easily end up assuming that axioms, laws, postulates, theories, and systems are immutable—

even though common sense and history say otherwise. This fundamental confusion seems not to disturb enough people: when I was a consultant to the Connecticut performance assessment project in math and science, I proposed the following challenge as a worthy task for assessing understanding of geometry:

> Two mathematicians had a debate. The first said that the postulates of geometry are like the rules of games: a system of rules and a mental model for thinking about space, but not "real." The second disagreed, saying that geometry is more like the features of the world and the textbook therefore more like a roadmap, a guidebook to what space "really" is like. Professor A's views seem to imply that geometry was *invented by mathematicians*, while Professor B seems to suggest that geometry was *discovered* (in the same way that America and the roundness of the earth were discovered). Who do you think was more correct and why? You work for a national student-focused magazine. Your editor wants you to come up with a lively article on the debate, giving examples most supportive of each side, interviews with audience members on their different reactions, and reasons why a reader of the magazine should care about the debate.

Only Steve Leinwand, the head of secondary mathematics education for Connecticut (and a delightfully unorthodox thinker in his own right), grasped both the importance of my prompt and the alarming implication of its rejection by all the teachers.[24] Can a student truly be said to *understand* geometry who is unaware of the limits of the system or the reasons behind the extraordinary move away from thinking of geometry as "real" to thinking of it as axiomatic? Even our best math students are unaware that non-Euclidean geometries can be proven to be as logically sound as Euclid's; fewer still are helped to understand the sea change in our thinking about what knowledge is that resulted when geometry could no longer be viewed as a system of truths.

The second result of our failure to teach and test for the history of ideas is that it does lasting harm to intellectual courage

in all but our feistiest students. Students never grasp that "knowledge" is the product of someone's "thinking"—thinking that was as lively, unfinished, and (sometimes) muddled as their own. One major reason for the intellectual poverty in this country is that most students become convinced either that they are incapable of being intellectual or that they are uninterested in being intellectual, thinking that it involves only the arcane expertise of a narrowly framed and inert subject.

Some practical assessment implications? First, we should require that students keep notebooks of reflections on coursework, their increasing knowledge, and important changes of mind about that knowledge. Second, we should assess this work as part of the grade. I did so for many years and found the notebooks to be the most important and revealing aspect of the students' work. I also learned a lot about how their thinking evolved in a way that improved the courses I taught. Third, the most technical of trainings should ask students to do critical research into the origins of the ideas being learned, so that students can gain greater perspective on their work. To fail to do this, whether out of habit or rationalization that there is no time for such reflection, is to risk producing a thoughtless batch of students.

Postulate 9: We should assess students' intellectual honesty and other habits of mind.

To worry about whether understanding is "thoughtful" or "thoughtless" is ultimately to concern ourselves with whether students are honest or dishonest about what they know and how they have come to know it.

I am not referring to the obviously heinous crime of cheating—something we know to be all too common. Rather, I am talking about the moral obligation of the student to emulate Socrates' trademark: his cheerful admission of ignorance. Alas, our students rarely do admit their ignorance. Thus one of our primary tasks should be to elicit (and not penalize) the admission. But the student's willingness to risk the admission depends upon *our* willingness. It is only after both teacher and student have admitted ignorance, as the Meno dialogue reminds us, that mutual inquiry and dialogue become possible; only then are we placed on equal moral footing as thinkers. Unfortunately, our inclination to "pro-

fess" is always in danger of closing the doors through which our students can enter the liberal conversation without excessive self-deprecation: so many of our students preface a wonderful idea by saying, "I know this sounds stupid, but. . . ."

Let our assessments therefore routinely encourage students to distinguish between what they do and do not know with conviction. Let us design scoring systems for papers that heavily penalize mere slickness and feigned control over a complex subject and greatly reward honest admissions of ignorance or confusion. And let us ask students to write a paper in which they critique the previous one they wrote.

Intellectual honesty is just one aspect of self-knowledge. Another important aspect is the absence of self-deception. This too can be furthered through assessment. One of my favorite notions along these lines was something the atomic physicist Leo Slizard is reputed to have said about how to assess doctoral candidates. He argued that students should be assessed on how precisely and well they know their strengths and limitations and felt that it was a mistake to err greatly in *either* direction. I am not arguing for teachers to become counselors or depth psychologists; I *am* arguing for their responsibility in improving the students' ability to self-assess in the deepest sense. We must ensure that students have neither excessive nor deficient pride in their work, either of which closes off further intellectual challenges and rewards.

The inherent danger of all scholarship is not so much error as blind spots in our knowledge—blind spots hidden by the increasingly narrowed focus of our work and the isolation that can then breed worse: arrogance. Excessive pride leads us not only to ignore or paper over our doubts but more subtly to be deceived about the uniqueness and worth of our ideas. We forget that it was a conversation with others in the coffee shop or an article in a recent journal that sparked the idea. A few collaborative assessment tasks, with some required reflection on every student's part about the roles of each contributor, would provide useful perspective for everyone. Similarly, students in the ARTS PROPEL program in Pittsburgh were required to thoroughly document the "biography" of a work—from the germ of the idea written on the back of an envelope, to peer criticisms, to the final draft—so as to see how ideas unfold and are influenced by others.

Given the importance of collaboration, we should also assess class discussions more thoroughly than we do. We again fail to assess what we value when we make it possible for students to learn everything that we deem necessary just by listening to us and doing the reading. Over the years, I have developed materials for assessment (and self-assessment) of class discussions, one example of which is found in Exhibit 2.2. (Each student fills out an evaluation after each major discussion, and the teacher fills out a simplified version on a weekly basis for each student.)

Which brings us back to Socrates, the discussant as teacher. What casual readers of Plato—(and even some overly analytic philosophers) always fail to grasp is that the dialogues invariably are about the role that character plays in intellectual development; they are never about mere "theories" of virtue, knowledge, or piety. The twists and turns of dialogue, the sparring with Sophists or young know-it-alls ultimately are meant to show that character flaws, not cognitive defects, impede the quest for a lifelong education. It is our *attitude* toward knowledge that ultimately determines whether we become wise (as opposed to merely learned.)

As Socrates repeatedly reminds us, we must love wisdom enough to question our knowledge—even our pet ideas, if need be. By extension, the more we gain confidence in our ideas, the more we must become vigilant about finding knowledge in unexpected places: we must imagine that those who seem incapable of wisdom might teach us something (as our students often do).

It is therefore not a canon—of ideas or books—that defines the liberal arts, but a set of very hard-won virtues. Like all sophisticated dispositions, these liberal habits are typically revealed only when they are challenged. It is only when peer pressure is greatest—be it in the classroom with students or at conferences with our peers—that we learn who has the power to keep questions alive. (Remember Gadamer's reminder that it is not ignorance but dominant opinion that is the enemy of thoughtfulness.) The liberal arts, properly speaking, do not *make* you free; they *keep* you free. Wisdom—as Socrates knew—reveals itself when persistent inquiry is threatened: externally by custom and such remarks as, "Oh, *everyone* knows . . . ," and internally by the tendency to rationalize our own habits, beliefs, and fears.

How much do students really love to learn, to persist, to

Exhibit 2.2. Discussion Rating Scales.

How did you feel about today's discussion?

Class's treatment of issues

superficial	1	2	3	4	5	thorough and deep

Helpfulness of discussion to your understanding

low	1	2	3	4	5	high

Your own level of engagement

low	1	2	3	4	5	high

The class's overall level of engagement

low	1	2	3	4	5	high

Quality of your own participation

poor	1	2	3	4	5	excellent

Quantity of your spoken remarks relative to your normal performance

low	1	2	3	4	5	high

Degree of your own understanding of material

limited	1	2	3	4	5	full

Facilitator's Success

too much input	1	2	3	4	5	too little input
too much control	1	2	3	4	5	too little control
great respect for others	1	2	3	4	5	too little respect for others

Comments

passionately attack a problem or task? How willing are they, like many of the great Native American potters of New Mexico, to watch some of their prized ideas explode and to start anew? How willing are they to go beyond being merely dutiful or long-winded? Let us

assess such things, just as good coaches do when they bench the talented player who "dogs" it or when they thrust the novice into the lineup because the novice's pluck impresses them more than the lack of experience.

We must make habits of mind—the intellectual virtues—central to our assessment. It is to our detriment and the detriment of the liberal arts that we feel squeamish about saying and doing so. The Scottish are not so squeamish: one report card I saw assessed students in terms of achievement, perseverance, and *flair!*[25] For that matter, most references requested of teachers by college admissions offices involve an assessment of those virtues. Consider, for example, the "universal" form used by a large group of private colleges (see Exhibit 2.3).

Let us thus routinely "test" students in the same way that a mountain "tests" the climber—through challenges designed to evoke the proper virtues, if they are present. And if they are not present, the quality of the resultant work should seem so inadequate to the *student* that little need be said in the way of evaluative feedback.

Let our assessments be built upon that age-old distinction between wisdom and knowledge, then. Too subjective? Unfair? Not to those who have the master's eyes, ears, and sense of smell—who have *tact*, in the old and unfortunately lost sense of that word. For these intellectual traits are as tangible as any fact to the true mentor, and they are more important to the student's welfare in the long run. It is not the student's errors that matter, but the student's responses to error; it is not mastery of a simplistic task that impresses, but the student's risk taking with the inherently complex; it is not thoroughness in a novice's work that reveals understanding, but full awareness of the dilemmas, compromises, and uncertainties lurking under the arguments he or she is willing to tentatively stand on.

If our typical testing encourages smug or thoughtless mastery—and it does—we undermine the liberal arts. If our assessment systems induce timidity, cockiness, or crass calculations about grades and the relevance of each assignment, we undermine the liberal arts. If our assessments value correctness more than insight and honesty, we undermine the liberal arts. If our assessments value ease of scoring over the important task of revealing to students the

Please feel free to write whatever you think is important about the applicant, including a description of academic and personal characteristics. We are particularly interested in the candidate's intellectual purpose, motivation, relative maturity, integrity, independence, originality, leadership potential, capacity for growth, special talents, and enthusiasm. We welcome information that will help us to differentiate this student from others.

Ratings

No basis	*Academic skills and potential*	*Below average*	*Average*	*Good*	*Very good*	*One of the top few encountered in my career*
	Creative, original thought					
	Motivation					
	Independence, initiative					
	Intellectual ability					
	Academic achievement					
	Written expression of ideas					
	Effective class discussion					
	Disciplined work habits					
	Potential for growth					
	Summary Evaluation					

BEST COPY AVAILABLE

errors or tasks that matter most, we undermine the liberal arts. Let us ensure, above all else, that our tests do just what Socrates' tests were meant to do: help us distinguish genuine from sham authority, the sophists from the wise. Then we will have assessments that are worthy of our aims.

Notes

1. This chapter is a substantial revision of a speech given to the American Association of Higher Education convention in Washington, D C., in June of 1991.
2. *Plato: Laches, Protagoras, Meno, Euthydemus,* W.R.M. Lamb, trans. (Cambridge, Mass.: Harvard University Press, 1962).
3. This is made quite clear in other Platonic dialogues. Theaetetus, for example, falls into some of the same self-inflicted logical errors made by Meno, but he is more aware of and honest about his mistakes, prompting Socrates to commend him on his forthrightness and adaptation to the conversation. See *Plato,* W.R.M. Lamb, trans.
4. G. Wiggins, "A True Test: Toward More Authentic and Equitable Assessment," *Phi Delta Kappan* 70 (1989a): 703-713.
5. A. H. Shoenfeld, "Problem Solving in Context(s)," in R. Charles and E. Silver, eds., *The Teaching and Assessing of Mathematical Problem Solving* (Reston, Va.: National Council of Teachers of Mathematics/Erlbaum, 1988), p. 84.
6. Many critics of alternative assessment are also convinced that the critics of conventional testing are against all testing. See J. Cizek, "Confusion Effusion: A Rejoinder to Wiggins," *Phi Delta Kappan* 73 (1991): 150-153. I confess that the conservative has a valid point here: many critics of standardized tests reveal themselves to be against all formal evaluation in reference to standards. That is a mistake, and it is one reason that alternative school people end up shooting themselves in the foot: they sometimes produce free spirits who are not very capable.
7. T. S. Kuhn, *The Structure of Scientific Revolutions,* 2nd ed. (Chicago: University of Chicago Press, 1970), pp. 165-166.

8. *Meaning* and *significance* are *not* synonyms, as C. K. Ogden and I. A. Richards's classic treatise on the meaning of meaning reminds us. (See C. K. Ogden and I. A. Richards, *The Meaning of Meaning*, 5th ed. [Orlando, Fla.: Harcourt Brace Jovanovich, 1938].) Meanings, they argued, are "objective," because they are derivable from the text or facts at hand. Significance, however, is more personal—projective and contextual. Whether or not this view is sound for literary criticism (consider the arguments between deconstructionists and hermeneuticists, for instance), it has a commonsense appeal. Thus the Pythagorean theorem may not have much significance to the average fifteen-year-old, but a great deal of important meaning and connection can be found in reference to it.
9. M. Foucault, *Discipline and Punish*. (New York: Vintage Books, 1977), pp. 181–184.
10. See Chapters Four and Five for more on the relationship between assessment, incentives, and tact.
11. See A. Collins, J. S. Brown, and S. E. Newman, "Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing, and Mathematics," in L. B. Resnick, ed., *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser* (Hillsdale, N.J.: Erlbaum, 1989), on cognitive apprenticeship, and H. Gardner, *The Unschooled Mind: How Children Think and How Schools Should Teach* (New York: Basic Books, 1991).
12. From International Baccalaureate Examination Office, Extended Essay Guidelines (Cardiff, Wales: International Baccalaureate Examination Office, 1991).
13. See G. Wiggins, "Creating Tests Worth Taking," *Educational Leadership* 49 (1992): 26–33, and Center on Learning, Assessment, and School Structure, *Standards, Not Standardization*, Vol. 3: *Rethinking Student Assessment* (Geneseo, N.Y.: Center on Learning, Assessment, and School Structure, 1993), for more on the design of performance tasks.
14. From the 1983 Advanced Placement exam in U.S. history.
15. A vivid picture of such self-assessment in action can be found in G. Grant and W. Kohli, "Assessing Student Performance," in G. Grant and Associates, *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education* (San Francisco: Jossey-Bass, 1979). This comprehensive study of college competency-based programs is essential reading for K-12 educators, especially those involved in outcomes-based education programs.

16. See P. Elbow, *Writing with Power: Techniques for Mastering the Writing Process* (New York: Oxford University Press, 1981).
17. No one has written more compellingly and informatively on this subject than Elbow in *Writing with Power*. See Chapter Six where I discuss his ideas further.
18. H.-G. Gadamer, *Truth and Method* (New York: Crossroad, 1982).
19. See G. Ryle, *The Concept of Mind* (London: Hutchinson House, 1949).
20. J. Bruner, *The Process of Education* (Cambridge, Mass.: Harvard University Press, 1960/1977).
21. This line of argument has profound consequences for scoring rubrics. There is a need to ensure that the scoring rubric is written "backwards" from the deepest and most penetrating understanding, therefore. The case for "developmental" and "progress" (longitudinal) assessments is discussed in Chapter Five. Such a system is now in place in Great Britain.
22. See G. Wiggins, "Creating Tests Worth Taking."
23. See J. Schwab, *Science, Curriculum, and Liberal Education* (Chicago: University of Chicago Press, 1978).
24. For readers interested in understanding the historical importance of this debate, see M. Kline, *Mathematics in Western Culture* (New York: Oxford University Press, 1953), and M. Kline, *Mathematics: The Loss of Certainty* (New York: Oxford University Press, 1980).
25. D. Archbald and F. Newmann, *Beyond Standardized Testing: Authentic Academic Achievement in the Secondary School* (Reston, Va.: NASSP Publications, 1988).

3 The Morality of Test Security

It is so common that we barely give it a second thought: the tests that we and others design to evaluate the success of student learning invariably depend upon secrecy.[1] Secrecy as to the questions that will be asked. Secrecy as to how the questions will be chosen. Secrecy as to how the results will be scored. Sometimes secrecy as to when we will be tested. Secrecy as to what the scores mean (if we are not given back our tests and an answer key). Secrecy as to how the results will be used. What a paradoxical affair! Our aim is to educate, to prepare, to enlighten, yet our habits of testing are built upon procedures that continually keep students in the dark—procedures with roots in premodern traditions of legal proceedings and religious inquisitions.

As with all deep-seated habits, we have lost sight of the questionable aspects of the practice beneath our rationalizations. "Surely we want the student to be able to grapple with the unexpected. . . . How could we *possibly* give the student access to the questions in advance? . . . There is no way to obtain validity without test security. . . . Hasn't testing *always* been done this way? . . . Isn't our secure testing system more 'fair,' since it ensures that everyone is judged in the same way?" The unthinking character of these responses would be plain if not for the ubiquity and seductive ease of the use of secrecy. We need only look at technical guides on how to enforce test security to become alert to the moral and intellectual dangers inherent in the practice. For example, in an old

132

BEST COPY AVAILABLE

textbook on testing, we are advised to "deny the examinee information concerning the rightness or wrongness of his response" as a way of keeping the test secure for other potential test takers—though the harm to the learner is inescapable and overt in such a procedure.[2]

Why would we take for granted that students do not have a right to full knowledge and justification of the form and content of each test and the standards by which their work will be judged? The student's (and often the teacher's) future is at stake, yet neither has an opportunity to question the aptness or adequacy of the test, the keying of the answers, or the scoring of the answers. Why would we assume that any test designer—be it a company or a classroom teacher—has a prior right to keep such information from test takers (and often test users)? Why would we assume, contrary to all accepted guidelines of experimental research, that test companies (and teachers) need not publish their tests and results after the fact for scrutiny by independent experts as well as the test taker? Maybe the better advice to test makers is that offered twenty years ago by performance assessment researchers Robert Fitzpatrick and Edward Morrison: "The best solution to the problem of test security is to keep no secrets."[3]

Whatever the technical reasons for test security, it clearly does unintended harm to students. Steady doses of such secrecy may well beget a lasting form of student furtiveness in response. Students learn to fear admitting ignorance or being creative. Questionable or imaginative responses, rather than being valued by students as the building blocks of thoughtful understanding, are viewed nervously as potential mistakes. The aim becomes not personal knowledge but figuring out "what they want"—better safe than sorry, though the loss of intellectual autonomy and honesty may be deeply regretted later in life. If by *character* we mean intellectual courage and moral integrity, then character may well be threatened by tests that are perpetually secret. The legacy of cramming, cheating, and "teaching to the test" may be imitative responses on the part of students and teachers to a sanctioned deceptiveness.

Adults have lost their empathy here. We no longer feel the useless anxiety the student feels in a world of tests that are valid only because of the prior (and often subsequent) secrecy of the instru-

133

ment. Though we at some level "know" that risk taking and high-level performance do not emerge in a climate of secrecy, we fail to see how our test rituals sanctify such a stance. Yet if we think of public policy or personal histories, the danger becomes clear: "Secrecy can debilitate judgment whenever it shuts out criticism and feedback, leading people to become mired down in stereotyped, unexamined and often erroneous beliefs and ways of thinking."[4]

In this chapter, we will examine the different kinds of secrecy found in testing situations everywhere. The case will be made that, while each form of secrecy can be appropriate in specific contexts, there is a need for much greater clarity as to the limits of each. We have a duty to ask, Should the assessor be assumed to have the right to employ such secrecy, irrespective of its impact on students, teachers, and schools? *Or is the* tester *obligated to justify the practice or its extent?*

The euphemism "secure test" hides the fact that an always morally questionable practice lies at the heart of testing as we know it. By using the word *security*, we imply that we have a property that needs to be kept safely in our possession, as a fundamental right. But is the test maker's property inherently more important than the test taker's right to openness and due process in assessment? After all, the "secure" test reflects a form of unilaterally exercised power that cannot be examined or easily contested by the test taker. The use of secrecy is more than a matter of technical tactics. Casting the matter in moral and legal language alerts us to the problems *inherent* in the practice. Due process is threatened when *any* "judge" does his or her work in secret. Imagine, for example, the harm to our political system and to citizen insight into and faith in our legal principles if Supreme Court judges did not have to publish opinions to support their decisions. (Consider how much better tests might be if every test had to be justified in writing in the same way that court decisions are made.[5])

There may well be times, in the classroom as well as at the state and national levels, when test security can be justified. What can no longer be justified, however, is the unspoken view that psychometricians and teachers have the right to make such secrecy a cornerstone of test methodology.

BEST COPY AVAILABLE

The Emperor Revisited

The practice of test security is so much a part of the educational landscape that a few stories and vignettes may be required to cast the problem in a fresh, revealing light. Because secure testing is omnipresent, we may no longer see how our judgment has become impoverished or our conduct stereotyped. We might do well, therefore, to consider our unthinking use of secret tests as a modern version of "The Emperor's New Clothes."

You no doubt recall the story. Rascals posing as tailors "wove" a suit of the finest "cloth" for the king, earning riches by fashioning an illusion—an illusion not only about the garment itself but about their skill in serving the king. The king's nakedness there to be seen by all, remained unseen. A sham that should have been obvious worked precisely because of the tailors' warning: only boorish, crude folks would fail to recognize the quality of the incredibly fine yarn. And so the townspeople rationalized their perceptions of nakedness and their secret doubt; they, like the king's retinue (who feared for their honor), praised the king as he paraded in his "finery." The king too, "knowing" that *he* was not a commoner, was sucked into the self-deception.

It was an innocent child, unaware of the "secret" and of the need to maintain secrecy, who exposed the hoax. "But he has nothing on!" exclaimed the child. But that is not the end of the story—and it is the ending that shows how harmful to judgment secrecy and the uncertainty about what we "should know" can be. The townspeople did not immediately come to their senses; the elders initially dismissed the remark of the young "innocent." Eventually, though, the child's words were repeated enough that their truth cut through the self-deception and doubt. But the Emperor, thinking that the now-skeptical townspeople must be right, "thought to himself, 'I must not stop or it will spoil the procession.' So he marched on even more proudly than before, and the courtiers continued to carry a train that was not there at all."[6]

The tale is instructive about current testing policy on many levels. We *still* do not want to spoil the procession. Testing increases, though few useful results emerge from the investment. We are *still* dismissing the remarks of the "innocents." We do not see

through the eyes of students as they prepare for and take the tests that we buy and realize how debilitating those tests are to intellectual engagement, courage, and imagination. Nor do we see through the eyes of employers, teachers, and administrators and realize how rarely they study test results to *understand* applicants' and students' abilities or the meaning of their errors. We are so self-deceived that we often comment, when discussing test reform, "But *we* made it through the system, didn't we? . . ."—as if these high-stakes secure tests were nothing more than the harmless indignities of a freshman initiation of years gone by.

Commercial test makers literally profit from the illusion that, like the clothes made from the tailor's yarn, all "fine" tests must be built with a specialist's mysterious skill. Testing, rather than being a common practice of assessing student performance on the tasks we value, becomes an arcane science that is entrusted—and apparently only entrustable—to statisticians.[7] Critics of such tests fear looking like the crude folks that the tailors warn people their critics will be; wary practitioners are routinely made to feel ignorant of the true "finery" in test validity and reliability.

The harm of any long-standing secret is its power to cause self-deception—as the story shows so clearly. The utter simplicity of the test items is like the king's nakedness: *so* obvious as to make one feel certain that some complex and unknown set of standards must render the seemingly nonexistent test "garment" substantive. Like the townspeople in the story, everyone—from teachers to superintendents—ends up talking as if the real capacities we value were being directly observed in detail (instead of by means of one or two proxy questions with simplistic answers, as is always the case). The supposed necessity of secure test questions eventually becomes "obvious" to everyone—as if all the important tests and criteria in life (for getting employed and getting a raise, writing a successful dissertation for a doctorate, obtaining a driver's license, winning the big game, or submitting a winning engineering or graphics-design bid) also involved secret tasks, criteria, and standards. The mystery of test-maker authority ensures that private doubts remain inchoate; highfalutin (but hazily understood) technical language comes to dominate our conversation.[8]

The inevitable happens: teachers imitate the form of secure

standardized tests, forgetting (or never realizing) that the technology is limited in its capacity to assess what we value. Even teachers who talk of the foolishness or harm of secure simplistic tests end up employing their own versions of them—the true sign of the tests' mythic rather than rational power. The arcane (but misunderstood) procedures of the "tailors" take root and are poorly mimicked by the uninitiated: any inspection of local tests reveals that almost all of them are of questionable validity and reliability.[9] Grades that teachers then give become increasingly less justifiable (even as they become more precise) as so-called objective tests proliferate. And the school transcript remains as unreliable as it ever was. The call for sounder standardized tests then naturally increases from the outside critics, and the vicious circle continues.[10] Bring in the tailors! Let the king march more proudly! But pity the poor student. For, contrary to the story, the child's voice—common sense—remains unheard or unheeded still.

The One-Sidedness of Secrecy

Secrecy is not inherently wrong, as Harvard University professor of philosophy Sissela Bok stresses in her fine treatise on the subject. But because there are moral consequences at stake in any use of secrecy, some overarching principles are necessary both to inform and safeguard our judgment and to inform technical policy and practice.

In professional, personal, and social matters, mutual respect must underlie any justified use of secrecy. Because a student starts with limited rights in any formal educational relationship, and those rights are further restricted by the traditional testing context, test makers and users are obligated to be more respectful of the student's position than they tend to be. This obligation is easier to see if we think of secrecy in relation to adult citizens or consumers, as Bok suggests: "No just society would . . . allocate controls so unequally. This is not to say that some people might not be granted limited powers for certain purposes . . . but they would have to advance reasons sufficient to overcome the initial presumption favoring equality."[11] The right of the persons affected to control what is kept secret becomes more compelling at the level of social insti-

BEST COPY AVAILABLE

tutions: "When power is joined to secrecy, the danger of spread and abuse . . . increases. In all such cases the presumption shifts [away from the assumption of a right to secrecy]. When those who exercise power . . . claim control over secrecy and openness, it is up to them to show why giving them such control is necessary and what kinds of safeguards they propose. . . . Even where persuasive reasons for collective practices of secrecy can be stated, accountability is indispensable."[12]

Bok does not address the issue of secrecy in testing in her book—a surprise, when one thinks about the inequities of power at stake in testing. She does, however, discuss the danger and moral harm of secretly intrusive psychological research, particularly research based on deceit by the researcher. And she approvingly quotes Margaret Mead's criticism of such research—criticism that was in part based on the fact that such methods "damage science by cutting short methods that would responsibly enhance, rather than destroy, human trust."[13]

No person or group should thus be assumed to have unilateral power to control what is kept secret. But this principle has always been difficult to honor when dealing with children (or others who seem "inferior" to us in terms of rights). We do, however, now take the right to be protected from secrecy for granted when dealing with judicial inquiries and the rights of the accused, and children can now sue their parents or be protected from abusive parents by the state. Due process certainly requires that secrecy be minimized in these arenas. But it was not always so. The insistence on mutual respect and openness in formal inquiries is recent, as Foucault notes. His history of legal investigation and examination in seventeenth-century France is a reminder of how difficult moral equality is to uphold in practice:

> In France, as in most European countries, with the notable exception of England, the entire criminal procedure, right up to the sentence, remained secret: that is to say, opaque, not only to the public but the accused himself. . . . [K]nowledge was the absolute privilege of the prosecution. The preliminary investigation was carried out as "diligently and secretly as may be," as the edict of 1498 put it. According to the ordinance of 1670, which confirmed . . . and reinforced the severity of the preceding period, it was impossible for the accused to have access to the documents of the case . . . impossible to know the nature of the evidence . . . impossible to make use of the documents in the proof.[14]

The combination of secrecy and unilateral respect could justify even deceit in the judge's conduct—a result that we would now find morally repugnant: "The magistrate, for his part, had the right to accept anonymous denunciations, to conceal from the accused the nature of the action, to question him with a view to catching him out, to use insinuations. . . . The secret and written form of the procedure reflects the principle that . . . establishment of truth was the absolute right and the exclusive power of the sovereign and his judges."[15]

It is not pushing the argument too much to ask the reader to reread these two passages and think of the student as the accused and the test maker as the judge. ("Knowledge as the privilege of the assessor," "impossible to know the nature of the evidence," "to question him with a view to catching him out" [as with the use of distracters in tests], and so on.) At the heart of Foucault's analysis is the view that such one-sided practices have been common to all the areas in which we seek to "discipline" humankind (law, military, education, and psychiatry). Foucault explicitly links the judicial and educational "examination," which "combines the technique of an observing hierarchy and those of a normalizing judgment." While he argues that "investigation" has become modernized through the methods of science, the "examination is still caught up in disciplinary technology."[16] Sue E. Berryman at Teachers College, Columbia University, notes, for example, that tests are "obscure, opaque, inaccessible"; "these tests and their results carry no intuitive meaning to anyone besides educators. . . . They thus fail to measure objectives that parties with interests in the outcomes of our educational system can understand, 'see,' and debate."[17]

The legacy of secrecy in student testing is of course long-

BEST COPY AVAILABLE

standing, with its roots in the world of autocratic religious power and the hierarchical, guild mentality of the Middle Ages—a world filled with many adult "secret societies."[18] The secret test as we know it came from this world, where professors saw themselves as members of a guild, where status transcended rights and was granted by "degrees" to only a few, where the assessor had the historically unquestionable right to demand (orthodox) answers of the test taker without justifying either the questions asked or the evaluation of the answers. Whatever modern justifications are made for test security on validity grounds, there is little doubt that the roots of the practice derive from this earlier assumed right to keep vital information from the examinee.

The modern world's unwillingness to tolerate such one-sided secrecy becomes clear when we examine the policies that have arisen around the use of secure tests in the hiring and promotion of *adults*. Irrespective of the wishes or habitual practices of testing specialists, the courts have been quite clear that the assessee has a right to more information (and more stringent, user-friendly tests of validity) than testers have historically wanted to provide (as we saw in Chapter One).[19]

It seems to me no coincidence that we continually and unthinkingly retain the use of blanket secrecy in dealing with minors. To put it charitably, we know better than they what is in their interest. To put it somewhat cynically, children will never likely be protected from excessive secrecy (or any other practice that depends upon moral inequality), because they will always be without adequate political clout to ensure that their moral equality is codified into law. Thus Walt Haney, Boston College researcher and co-head of the National Center for Testing, who has been an expert witness in the long-standing lawsuit over whether test companies are obligated to release individual completed and scored tests (as opposed to just scores), told me that in his opinion the original test disclosure law would not have been passed if the major class of litigants had been schoolchildren, not medical students.[20]

The Formal Positions of the Professions on Test Ethics

We look in vain for adequate guidance from the affected professions on the ethics of secrecy in student assessment. As we saw in Chapter

BEST COPY AVAILABLE

One, the APA Standards for educational and psychological testing were designed to provide criteria not only for the evaluation of tests but for "testing practices, and the effects of test use."[21] These APA Standards note that the "interests of the various parties in the testing process are usually, but not always, congruent." The only example given of noncongruent interests has to do with testing for admission to highly selective jobs, schools, or programs. But why wouldn't many teachers and older students cite secure tests as not congruent with their interests or specific programs? Secrecy is certainly not in the interest of a would-be performer who is going to be judged.

In neither the technical nor the ethical standards do we find discussion of the test taker's rights with respect to adequate advance knowledge and preparation. (In fact, many of the standards under both test administration and test-taker rights have to do with the maintenance of test security!) And there is no discussion of what role, if any, student test takers and teacher test users should have concerning a test's validity for their particular context. The bulk of the standards deal with the test taker's rights of confidentiality and the harm of unobtrusive testing—that is, testing conducted without the test taker's knowledge: the issue of informed consent, more generally construed.[22] The only standard relevant to test security is Standard 16.2, which states that "test users should provide test takers . . . with an appropriate explanation of test results and recommendations made on the basis of test results in a form they can understand."

This "form they can understand" need not include, however, direct access to the answered and scored test. An intriguing historical change of mind that relates to access to one's scored test has in fact occurred in the APA Standards for ethics in research. In a recent paper, Haney and Madaus ruefully point out that a key clause of the 1977 ethics guidelines was ***deleted*** in the most recent (1981) edition.[23] Beyond the test taker's "right to know results" and "interpretations made" on the basis of results, the new version no longer contains the following clause from the earlier document: the test taker was deemed to have the right, "where appropriate, [to] the original data on which final judgments were made." It was this "right" that led to the test disclosure law in New York—a law

bitterly fought by test companies then and still a matter of litigation today (in a lawsuit fought on the grounds that the test company's copyright is more salient than the test taker's right to the completed test and answer key).

While ex post facto test disclosure has been debated in this way for years, only recently have arguments been made for greater openness *prior* to the test. Following up on an idea presented by Jerrold Zacharias years ago, Judah Schwartz of Harvard has been among the most vocal and eloquent in arguing that the complete item bank from which any test might be constructed also be made public.[24] Given that any test is a sample; and given that the item bank can be made large enough to encompass the domain of possible questions and thus prevent narrow cramming and invalid inferences about results, there is little reason to maintain perpetual security, and there are many good reasons for making that bank of questions open for public inspection. We could easily develop "large publicly available data bases of reviewed and signed problems in a wide variety of school subject areas. . . . School systems and state boards of education and other responsible authorities would use these data bases as the sources of the questions and problems in the accountability [systems]."[25]

Schwartz proposes that the state or district make clear which sections of the data base would be tapped and what the criteria would be for the selection of particular questions. The data base would be available in all "libraries and bookstores," with the clear advantage of such a "sunshine" procedure: "If the pool of problems from which examinations are composed is publicly available in advance, then any group in society that feels ill-served by the substance or language or context of a question can raise a prima facie objection in advance of the use of the question and not have to rely on the vagaries of statistical analyses to claim bias or prejudice after the fact."[26]

The Student's Perspective: Some Revealing Vignettes

We need to be reminded that there exist ways of examining that do not rely on total test secrecy. Indeed, in any system designed to produce high-quality performance by all participants, such secrecy

BEST COPY AVAILABLE

is antithetical to the success of the enterprise, as the military, the performing arts, and athletics reveal. And in the adult world of performance appraisal, we see that not only are such "tests" possible but that there exist laws and regulations that protect adult test takers from the kind of excessive secrecy we routinely foist upon students and teachers.

Consider first the following fanciful scenarios, all written from the test taker's point of view. They illustrate both how students experience the effects of unrelenting test security and how counterproductive secrecy is to would-be performers and performance systems.

Vignette 1

Imagine the student as the manager of a warehouse, where organization of incoming material and the ability to pull together future orders are paramount. But reflect on what the manager's job might be like if he or she did not know what kinds of materials would be arriving each day, the quantity of those materials, or the kinds of orders that would later have to be filled. This is the student's position in school.

Each day, new material arrives, sometimes too quickly for the student manager to organize it on the shelves. After a week of sorting through and ordering the contents, new and unexpected material arrives, compelling the student to completely rethink the system used in storing—with no assurance that the revised system will compensate for future, unknown deliveries.

After months of storing and organizing catch-as-catch-can, the student managers are warned by the central office that they must be prepared to correctly fill dozens of (unknown) orders on the spot, without access to the notes and resources that serve as their data base. The managers will be "tested" on their ability to predict the order and fill it from memory—not the more authentic and vital ability to plan for, refine, and routinely process a wide range of known and varying orders.

One can see from this vignette why cheating can easily become a way of life for those who cannot develop a trivia-oriented memory or psych-out-the-test-maker tricks on their own. The arbi-

triariness of the test and the inappropriate reliance on memory encourage the test taker to consider all possible means for mastering the challenge—especially since "rehearsal" has been ruled out by the design of the test. This focus is especially tragic since, as it is now, students are rarely taught to learn how to learn—that is, how to "manage" knowledge so as to effectively store and retrieve it for thoughtful, flexible use—nor are they assessed in such a way as to test their ability to manage available resources.

Vignette 2

To better appreciate how excessive secrecy in testing corrupts relationships, consider what our response as adults would be to a job evaluation process in which the employer could do what test makers routinely do: without our knowledge or consent, pick a few tasks from the hundreds we have learned and performed over the years, demand an on-the-spot performance, and assess that performance only. It is telling that, with adults, the practice would be regarded as unjust and likely illegal. Why does it not seem so when we deal with children?

A principal in the Buffalo, New York, area took me aside at a workshop to share with me a story that illustrates the potential for hypocrisy here. He recently negotiated a forward-looking and humane performance appraisal system with his teachers—an open system involving peer consultation and teacher self-assessment through a portfolio process. It occurred to the principal that the process was generalizable: such a system might well be extended to the performance appraisal of students. The teachers (especially the high school teachers) would hear none of it.

Vignette 3

Consider how the combination of security and proxy (indirect) secret items ultimately corrupts performance. Consider a performance system, such as a professional sport league. What would happen if baseball were played as usual all season long, but then the pennant races were decided by one-shot, secure tests designed by statisticians and resulting in one aggregate score? Thus on the last day of the

BEST COPY AVAILABLE

season, specially constructed secure tests would be given to each player, composed of static drills and game situations. The pennant races would be decided each year by a new test and its results. Who believes that this secrecy, so necessary to the test's validity, would not end up corrupting both the game and the players' motivation? (Note that the students' current situation is actually worse, because students are usually not allowed to "play the game" of knowledge but must endure syllabi composed of drills and contrived and discrete game situations ordered in "scope and sequence" fashion. Not ever learning the "game" of knowledge through actual use, students are even less likely to predict the kind of real-world challenges they will ultimately face.)

More than the first two vignettes, this third one reveals how unwittingly and inappropriately powerful the possessors of secret testing knowledge and criteria can be—even if their aim in life is to be helpful statisticians. The test designer here supposedly seeks only to design a valid sampling-test of all the important elements of performance as specified by others. The secrecy is "justified," because without it the test would be corrupted: coaches would "teach to the test"—in the bad sense of shortchanging the players' overall education so as to artificially raise the score and distort its meaning. Yet we easily see how such a system would corrupt coaching and the game itself anyway. Not only the players but also the coaches would be robbed of the capacity to concentrate on the time-consuming task of developing excellent play beyond drill in discrete skills—just as in the classroom, teachers never end up asking students to use knowledge in complex, integrated ways, given the demands of preparing for the multiple-choice test.[27]

Vignette 4

Imagine that student musicians have to wait until the day of the concert to know the music that they will be playing together. And suppose, in keeping with typical testing, that the test of musical skill is made using isolated musical "items" (bits of pieces, not a whole work). Assume too that students play their instruments through microphones connected to other rooms—microphones that allow judges to listen but prevent students from hearing themselves

play. And assume also that the scoring is norm-referenced: weeks later, the students receive a single score telling them where they stand relative to all the clarinet or trumpet players in the state and a computer printout summarizing the stylistic and technical areas they should work on. (From this information, teachers and students are meant to improve!)

This vignette reminds us that a one-shot, secure test cannot, *in principle*, reveal whether the student has and uses a repertoire for mastering complex performances. In what sense does the musician learn about where he or she stands from such a test? What would "normed scores" reveal about the quality of musicianship in the region? How can the musician *master* a secure performance played once? What have we gained by such artificial precision if neither students nor judges can equate the assessment results to the criterion performance and its standards?

While it is true that I have chosen the vignettes to illustrate what I trust are commonsense objections to test security, my aim is really to jog our thinking about this thoughtless habit of ours. Where, other than in schools, do officials *willingly* turn over accountability to outsiders who use secret "audit" procedures that are insensitive to local goals and practices? How can schools be held accountable by such tests if teachers and students do not know, have a say in, or trust the standards by which they will be judged? Where but in schools are teachers assumed to have the right to construct any old test they please, using items that require security before and after the test? Why should we *tolerate* such secrecy at the local level, given ample opportunities to design a set of tasks for reliability and the ability to demand multiday work on substantial projects that do not require security.

The Various Kinds of Test Secrecy

Given our unthinking reliance on secrecy in testing and the obvious harm that such secrecy inflicts on performance rehearsal and later improvement, it is worth our while to think through the issue of its necessity. Let us begin by considering the most basic question, then: In what sense *is* a "test" secret, and in what sense *must* a test be secret (in order that it not be compromised)? And whether or not

BEST COPY AVAILABLE

a test is compromised by being no longer secure, when is such secrecy unfair to the test taker or test user? Where do the test taker's rights become predominant, in other words, in a conflict of interest over security in testing? We will consider a variety of practices in which the testing that takes place in schools is predicated on a degree of secrecy about what will be assessed, what methods of assessment will be used, the value of those methods, what the results mean, and the value of those results.

Everyone has experienced the most obvious answer to the question of what is a secret in conventional testing. Each of us has entered many classrooms not knowing the specific questions that were going to be asked. Such secrecy is so common that it has a profound hold on our consciousness. "Secret up until the day it occurs"—that is what a test *is*, for most of us. The mythic power of this practice was made very clear in a story told to me by Sharon P., a student teacher who tried some novel approaches to instruction and assessment in her student-teaching practicum. She designed a culminating unit in her English course in which the students, in small groups, devised a selection of would-be questions for the final exam, according to certain criteria supplied by Sharon. The reaction of two teachers in the faculty room at her school summed up the irrational hold that the practice of test security has upon us: "But then the students will cheat!" they protested in unison. Nor were they convinced of the wisdom of such a policy even after a lengthy discussion.

If we examine more closely such situations and the issues raised, we see that there are diverse aspects of test secrecy that occur before, during, and after a test. I can think of eleven kinds of secrecy built into conventional testing. The first three occur *before* the test is taken:

- The *specific* questions that will be on the test are usually secret, known only to the test maker.
- A larger set of questions from which the later secure test will sample is often secret, although students sometimes have access to these (or to the data base described above).
- The timing of an upcoming test is sometimes secret, a common feature of the high school or college "quiz"—usually described

euphemistically by the teacher on the day in question as "unannounced."

The following kinds of secrecy may exist *during* the test:

- The test itself can be secret, in that the student is actually unaware that a test is taking place. (In the technical literature, this is called "unobtrusive" testing.)
- The scoring criteria and standards can be secret (when the questions are not multiple-choice or equally weighted in value), in that the student is unaware of the scoring criteria, descriptors for the range of answers, and the "anchor performances" being used to set standards.
- Whether or not one is on the right track in understanding a question or forming an approach to it is secret. The student cannot converse with the test maker or test administrator, in other words, to check on the meaning of a question or the aptness of an approach to answering it.
- The resources for constructing and double-checking possible answers are kept secret (to stretch the term a bit), in the sense that the real-world resources that would be appropriately available are explicitly kept from the student during the test. This type of security extends to the use of other people as resources, including not only students but any adults present.

The following kinds of secrecy refer to what is not revealed *after* the test:

- The meaning of the results may be kept secret, as when the tester does not release the actual test paper. Students are thus unable to confirm or critique their score or the test itself. As I mentioned in the previous chapter, undergraduates frequently do not receive back their blue books after an exam.
- Even if the students get their papers back, the results remain somewhat secret if students are not allowed to see the answer key or sets of exemplary test papers against which to compare their answers.
- The significance of the national or state multiple-choice test is

kept secret by virtue of its indirect nature. It almost never has obvious relevance to local curriculum or "face validity" to the student and teacher. And when technical validation for the test tasks and standards does exist, it is rarely meaningful to the student and the teacher.
- The value of the test for future learning and aspirations remains secret in testing that is one-shot, not scaled using longitudinal scoring criteria and standards, and composed of indirect items.

There are arguments for the use of each one of these secrecy practices in certain instances. What we seek, however, are some principles for judging the boundary lines in each case. In what contexts is each type of secrecy appropriate and in what contexts is it not? How might we better judge when any particular use of secrecy in assessment goes over the line and becomes unjustified?

To address these questions, two prior sets of questions must be kept in mind. The first set is technical and will be addressed further in Chapter Seven, in a reconsideration of validity. The questions in that set boil down to this: What are we *really* testing by relying on so much secrecy? In other words, leaving aside correlations with similar types of tests, can we really be said to be testing for scientific or historical understanding, for example, if the student has no prior knowledge of the question, no opportunity to use resources, and no opportunity to ask questions? A second set of questions is moral, and it is these questions that are the focus of our consideration in this chapter: What educational values are at stake with (and perhaps threatened by) the persistent use of such secrecy by adults with children? To what extent might even a form of secrecy in testing that is defensible on technical grounds *inherently* threaten teacher-student and student-subject relationships—with all that such a threat implies for the potential harm to the student's faith in the school, its agents, and its standards?

Pretest Secrecy

The technical argument for pretest secrecy is fairly straightforward. The validity of all short-answer tests (though not necessarily examinations and most authentic assessments) is typically compromised

BEST COPY AVAILABLE

BEST COPY AVAILABLE

if the test questions are known in advance. Why should this be so? For two different reasons. First, the student would then be in a position to "know" the correct answer without necessarily knowing why it is so, given the format. Knowing many questions in advance would enable the student to short-circuit the learning process. Students could simply memorize the questions and correct answers, perhaps even gaining the answers from someone else. But this result would render invalid the inference that the student who does well on the test "knows and understands" the subject matter tested!

Second, insight into the student's breadth of knowledge may be jeopardized. Since most tests involve a necessarily small sample of the total domain of a subject and what was taught, knowing the questions in advance can make the student wrongly appear to have a very expansive knowledge. (Put differently, in most testing situations, we want and expect to be able to generalize about the breadth of mastery beyond what was tested—just as we generalize to the whole nation from a sample of 1,100 in a Gallup poll.) Instead of making an effort to successfully gain control over the domain of all *possible* important tasks or questions, the student need only concentrate on the small sample of questions now known to be on the test. Note that a test of a few essay questions or performance tasks puts the teacher or test maker more at risk for such an invalid inference than does a multiple-choice test (if the essays are the entire test). Since the student need only prepare for one or two essay questions—versus the typical hundred or more questions on a multiple-choice test—the validity of any inference about student control of all the important topics may be severely compromised by advance knowledge of the questions.

This is a solvable problem, however, if the student is given access to a representative set of possible test questions and if we acknowledge that most tests overrely on low-level, information-focused questions. The model of distributing a set of possible questions in advance is, after all, a common feature of graduate and undergraduate education. The student studies all the questions carefully, in a methodical and comprehensive review of the priority questions of the course, and knows that some of those questions will compose all or most of the final examination. Some questions would admittedly go unasked if this were to be a common practice,

but we would do well to consider why we feel the need to ask those kinds of questions.

There are shades of gray as well. If we say that the questions should be secure, do we hold the same view about the *format* of the questions? At present, it is considered sound policy for teachers or testing companies and agencies to describe the types of questions to be asked (definition, true-false, and/or essay, for example) and to provide samples of each type of question for review. (It is just this kind of modest "declassifying" that enables SAT and ACT preparation services to function and thrive. If this sort of revelation is deemed a fair practice, why should we not offer some opportunities for students to practice some of the particular questions—for example, as a take-home part of an exam?) It would seem unfair on the face of it for the student not to know the various types of questions and their approximate importance in the whole test; it seems to violate basic principles of due process and equal access. (And a failure to ensure prior access to the test format might upset the technical validity of the results, if some students are familiar with the format and others are not.) Similarly, it would seem inappropriate for the student to not know the weight of a subscore or the criteria or standards being used when the test involves judgment-based scoring. How would I know how to apportion time or judge the length or depth of my response without such information?

But prior openness can be taken further, as Judah Schwartz's suggestion for a regional, state, or national data base makes clear. We could ask all classes of questions and have open testing, without compromising validity, if we defined *openness* as prior knowledge of the complete domain of all specific questions or tasks that we have in our collective possession. It would no longer be feasible to "cheat" on the validity issue, because it would not be practical to master each question in advance; students would still have to worry about the whole domain of knowledge that the test would tap. In many cases at the high school and college level, all possible test questions are known in advance course by course; what is "secret" is which few ones from the larger list will actually be used.

Indeed, if we thought of the "portfolio" as a set of products and performances meeting certain criteria with respect to range of genre, topic, or type of product, then it would be quite possible to

BEST COPY AVAILABLE

dispense with most security. Consider the performance-based course requirements shown in Exhibit 3.1, all of which could (and should) be known in advance and "taught to," without compromising the breadth of the domain, while making it possible for students to know their performance obligations.

We certainly should feel comfortable asking districts (or regions, using regional service agencies that exist in each state) to develop such a set of portfolio categories and/or a complete data base over the years, subject to some of the principles mentioned at the end of Chapter One (in the Assessment Bill of Rights).

One uncontroversial implication concerning this kind of openness of the domain and secrecy of the sample is that we would no doubt improve the quality of all tests and test questions. If all test makers, including teachers, had to publicly provide a larger pool from which the test would sample, the current preponderance of low-level recall questions having little to do with genuine competence in a field of study would be undercut. Thoughtful and deep understanding is simply not assessable in secure testing, and we will continue to send the message to teachers that simplistic recall or application, based on "coverage," is all that matters—until we change the policy of secrecy.

In short, we should not let the test maker off the hook with only the most superficial form of openness; we should require more. To be sure, it might engender more chaos and greater cost in testing if tests were not secure, but we should at least explore a cost-benefit analysis, given how clearly test openness is in the student's interest.

At the very least, we should require, as a criterion of good testing, that any educational test be as open *as is possible*. This would not take us very far in determining the limits, of course, but it would at least make the matter more explicit and therefore more subject to scrutiny and external review. It would compel test designers to think through the matter of the student's right to proper rehearsals more clearly, for example. And it would likely lead to tests that have varying components, some of which are appropriately known in advance and others which are not. (We see such a model used all the time in performance competitions such as Odyssey of the Mind, formal debates, and music competitions.)

It is of course the case, however, that in fields that involve the vagaries of human interaction, one cannot know precisely what

Exhibit 3.1. Major Tasks for a Global Studies Course

1. Design a tour of the world's most holy sites.
   - Include accurate maps.
   - Prepare a guidebook, with descriptions of local norms, customs, etiquette.
   - Analyze the most cost-effective route and means of transportation.
   - Write an interesting-to-students history of the sites.
   - Compile an annotated bibliography of recommended readings for other students.
2. Write an International Bill of Rights.
   - Refer to past attempts and their strengths and weaknesses: Helsinki Accords, Communist Manifesto, U.S. Bill of Rights, and so on.
   - Convince a diverse group of peers and adults to "sign on."
3. Write a policy analysis and background report on a Latin American country for the secretary of state.
   - What should be our short-term policy goals with that country?
   - What are that country's economic and political prospects?
4. Collect and analyze media reports from *other* countries on U.S policies in the Middle East.
   - Put together a "briefing book" of photocopied press clips for the president, with commentary on the accuracy of each story.
   - Videotape/audiotape a simulated newscast summarizing world reaction to a recent U.S. policy decision.
5. Compile an oral history on a topical but historically interesting issue:
   - Interview recent American immigrants.
   - Talk to veterans of Operation Desert Storm, Vietnam, and World War II about America's role as a police officer for world affairs.
6. Design a museum exhibit, using artifacts and facsimiles.
   - Exhibit links between a European country's geography and its economy.
   - Illustrate the local area's role in the industrial revolution.
   - Display patterns of modern emigration and their causes.
7. Write and deliver, on videotape, two speeches: the first, by the visiting head of an African country on the history of U.S.-Africa relations; the second, in response, by President Clinton's spokesperson.
8. Take part in a formal debate on a controversial issue of global significance—for example, aid to Russian republics or the U.S. role in the fall of communism.
9. Create a model United Nations (with groups of two or three representing each country) and enact a new Security Council resolution on terrorism.
10. Write a textbook chapter titled "The Primary Causes of Revolution: People, Ideas, Events, or Economic Conditions?" in which you weigh the various causes of revolution and reflect on whether the most important revolutions were "revolutionary" or "evolutionary."

BEST COPY AVAILABLE

the "test" will require: there are inherently unanticipatable aspects to real-world tests. The client does the unexpected; the other team seeks to outwit us; difficult weather conditions demand unusual piloting maneuvers, and so on. Some would use this as an excuse for retaining complete pretest security, but the real-world situation is quite different: the adult always has the opportunity to adjust, consult resources, ask questions, bring in others, or even seek a delay. And the diligent student can rehearse the task effectively even if in the "real" test the facts, problems, or contexts are varied. This rehearsal is what coaches of athletes, performing artists, and teachers of medicine and law help students do: know the difference between what is inherently anticipatable and what is not, work on what is, and train for the imaginable unanticipatable events. As General Schwarzkopf, the commander of Operation Desert Storm put it: hope for the best; plan for the worst.

Secrecy During the Test

The rationale for standardizing test conditions and materials is clear enough: fairness and validity require each student to have equal opportunity to answer a question correctly. If test makers do not standardize what resources can and cannot be used and develop a standard protocol for what test administrators can and cannot say during the test, then we run the risk of invalid inferences about performance and unreliable scores.

But that rationale hardly justifies the regular practice of forbidding almost all human interaction and the use of contextually appropriate resources—particularly if our aim is to make tests educative and authentic and to determine whether students can intelligently use resources.

One form of concurrent secrecy, unobtrusiveness (that is, testing without the student's knowledge), while seemingly desirable, is often harmful. "Informed consent"—the antithesis of unobtrusive testing—is an ethical principle we value; it is included in the APA ethical and technical guidelines as a desirable standard. Yet as I note in Chapter Four, the British have deliberately made such unobtrusiveness a goal of (imposed) testing as a way of ensuring that tests are more authentic and more seamlessly interwoven with daily instruction (and thus less likely to produce needless test anxiety). The goal of unobtrusiveness seems ironically problematic: as psychometricians Jay Millman and Jennifer Greene have pointed out, "An examinee's *maximum* performance might not occur when testing is unobtrusive."[29] But "unobtrusiveness" in the British sense involves avoiding an abrupt departure from the kinds of challenges and problems that face a student in the classroom. Surely that is a worthy aim, given that part of what is wrong with traditional models of testing is that the student has to gear up for methods and formats that have no carryover to authentic learning or real-world tests. On balance, then, it would seem that an attempt to minimize obtrusiveness while still honoring the basic principle of informed consent is the proper guiding principle.

With respect to the secrecy of resources during the test, the technical constraints requiring the standardizing of test conditions invariably seem to inappropriately outweigh the pedagogical ones. Keeping the essential tools of research and reference unavailable compromises the construct validity of a test, in my judgment. If the nature of the simplified items used in tests prevents the use of books or student notes—resources that would invariably be available during those times when the student is "tested" in the world of intellectual performance—then what are we testing? And what lessons about adult standards and authority are learned if resources are arbitrarily withheld—*arbitrarily* in the sense that the student knows full well that during all genuine work situations, resources are available. Here we see most clearly that the use of tests, combined with secrecy, causes a moral imbalance that needs to be closely monitored.

Shades of gray exist with respect to possible access to resources. One physics teacher I know allows each student to bring to the final exam an index card filled out as he or she sees fit. Not only can this device sharpen and focus prior study; it may provide the teacher with an additional assessment: Do the students know which facts, figures, and formulas are of most worth as resources? (An unintended outcome, of course, is that students learn to write in superhumanly small script!) Some state exams go halfway by "standardizing" the resources. New York's earth science exam, for

BEST COPY AVAILABLE

example, provides a booklet of tables and charts for students to use during the exam.

Whatever the technical merits in forbidding the use of resources, another pedagogically harmful aspect of the prohibition is the implicit lesson that what other people know that might be of help must remain secret as students wrestle with difficult challenges. Our testing, wittingly or not, daily makes clear the dysfunctional lesson that intellectual assessment must be conducted in silence and in isolation from others. To ensure the validity of the inference that a person's score is truly that individual's, such a practice may well be defensible in many instances. But another kind of validity may thereby be sacrificed. What are we actually finding out about our students if the most basic feature of adult work life—namely, that we can use all resources, including other people, as necessary to get the job done—is never assessed? We should not only not discourage the use of people resources; we should actively encourage it. We can, for example, build the effective seeking of advice into a test: many district writing assessments and performance science tests are now multiday, and they build into the process the possibility of collaboration after an initial rough draft in isolation.

There is potentially a deeper moral harm to the student when the only option for collaboration is an accomplice or cheater. Years ago Piaget went so far as to suggest that exam-oriented schooling, where educators "have found themselves obliged to shut the child up in work that is strictly individual," reinforces all the child's most self-centered instincts and "seems contrary to the most obvious requirements of intellectual and moral development."[30]

What, then, of the student's presumed right to ask questions of the assessor? Here is a problem of profound moral significance: How can we better honor the right of the questioned to question the questioner? But that right threatens traditional standardization and technical concerns for comparability. We know that tests that permit assessor/student interaction are easily corrupted unless there are strict protocols and guidelines by which adult judgment about intervention is informed. But we can turn the question around: How can the scores on inherently complex and ill-structured test questions be adequately reliable if, despite the invariable ambiguity of test questions, the student cannot clarify the questions or double-

check a proposed solution (as would be allowed in any realistic setting)?

All these situations call for test makers to develop more sophisticated questions and dynamic/interactive tasks, with a protocol for test administration that lays out in some detail the permissible and impermissible interactions. This has already been done in IQ testing and in a good deal of qualitative research (such as moral development interviews). And we should demand of test companies as well as teachers the inclusion of certain questions and tasks that permit the use of resources that are specified in advance (such as textbook, dictionary, calculator, and student notes)—questions that could be given out after a preliminary secure section, perhaps. (Students could procure the resources they brought [and/or resources provided by the tester] from another room during a break, after handing in their first efforts, in anticipation of the more open part of the test. Then there would be no compromise to the secure first part of the test.)

After-Test Secrecy

As for ex post facto security, there can be little justification for it, especially at the local level. While it is true that certain proprietary interests might be threatened by such a policy, I would argue that those concerns must be of secondary importance in almost all cases, because it is surely detrimental to the student's (and teacher's) education for the test papers to remain a secret. Arguments that such a policy imposes an undo hardship on test companies do not seem very compelling, when New York State, the International Baccalaureate, and the Advanced Placement programs make their tests and student papers public each year.

Being unable to inspect the test after the fact means that the value of the test—its formal validation—remains an inappropriate secret, especially when students and teachers never see nontechnical (and nonintimidating) evidence and argument for the validity of the test in the first place.[31] In the long run, this can become an issue of educational equity and access: the student test taker is prevented from gaining not only the incentives but the insights that come from tests that are authentic and thus educative—tests in which

BEST COPY AVAILABLE

tasks and scoring standards instruct the student about real-world tasks and standards.

Ex post facto test security may well contribute to the routine design of tests as isolated, one-event experiences—a design that is never in the student's interest. Under such conditions, the real meaning of each test score, and the significance of that score for the student, can therefore be said to be inappropriately "secret." The student receives what are, in effect, noncomparable scores over time (as opposed to navigating a system of recurring assessments, scored in terms of continuous progress in reference to known, stable standards; see Chapter Four). This practice is thus linked to the greatest moral harm in secure tests (especially those of the indirect or "proxy" kind): the inherent failure of such tests to provide powerful, immediate, and useful feedback to all students as to the nature of their errors, the validated importance of such errors, and diagnostic help in how the errors might be eradicated.

The usual criticisms of the multiple-choice test thus miss the most potentially damaging consequences of the format. It may well be that the items adequately and appropriately discriminate. But mustn't unending secrecy about those items induce a debilitating quest for mere correctness and cramming instead of evoking and soliciting the best of what the student can do? Unending test security easily breeds student insecurity (and teacher insecurity, when the multiple-choice test is externally designed). Students can never feel confident, going into a secret multiple-choice test, that this odd sample of test items will reflect their prior work, achievements, or intellectual strengths.

Piaget's Insight: The Development of Respect for Standards

Deprived of both prior and after-the-fact insight into test tasks and scores, students end up lacking more than a rationale for what they must undergo. They are deprived of the opportunity to fully understand, master, and respect the standards for judging intellectual excellence. Testing is thus seen as an arbitrary act, a game, something to be "psyched out," in the words of older students. The persistence of such a system sends a decidedly irrational and disrespectful message to students: we claim that our test questions are valid and our keyed answers are correct, though we need never justify or explain them to you (or even show you the answers later). Let us call this what it is: deeply disrespectful—and ultimately dysfunctional.

It was Piaget who astutely argued that mutual respect between adults and children has consequences for more than just the student's social development. He argued that *intellectual* standards (and a lived commitment to them) evolve out of a particular *moral* experience: a genuine, open, and ongoing dialogue among mutually respectful equals. Mere exposure to good assessing and "correct" answers does not produce in us a respect for the intellectual standards that underlie good answers. Autonomy develops only "when the mind regards as necessary an ideal that is independent of all external pressure . . . and appears only with reciprocity, when mutual respect is strong enough." Only in habitual discussion among relational equals arises the inward recognition of intellectual standards as "necessary to common search for truth."[32]

Developing the habit of feeling *self-obligated* to intellectual standards, Piaget argued, depends therefore on the experience of the constant need for justification that obliges everyone—in other words, the "rules" that require and enable *all* equal parties to agree with a result. But clearly such an experience is predicated on a mutual sharing of arguments, not on received opinions. To understand the value of the assessor's tasks, answers, and methods, I must be respected enough to have access to their rationale. The assessor must feel obliged to put me on an equal footing with respect to the *why*, not just the *what*, of that which I am asked to do.

With that mutual respect and willingness to justify comes a risk, of course: the assessor is now open to being proven mistaken or being seen as lacking an apt rationale. "One must place oneself at the child's level, and give him a feeling of equality by laying stress on one's own obligations and one's own deficiencies." The unwillingness to take such a risk is all too common, however—a fact constantly decried by Piaget: "It looks as though in many ways the adult did everything in his power . . . to strengthen egocentrism in its double aspect, intellectual and moral."[33]

The student's alienation from adult authorities who avoid

BEST COPY AVAILABLE

this risk is the inevitable result, and this alienation undermines more than just our trust in adults, according to Piaget; it also undermines our ability to grasp intellectual standards *as worthy in their own right*. An ongoing implicit appeal to Authority (for Authority's sake) by our adult judges, instead of an explicit appeal to shared principles, strengthens our "natural" egocentrism and keeps us from understanding what the expert knows. We come to assume that answers are right or wrong merely because some expert says so, not because there is verifiable evidence or justifiable argument supporting those answers.

We therefore may not ever come to see that the criteria of knowledge—*logic* in the broad sense (that is, rules of evidence and argument)—are verifiably worthy and that they merit being our own internal criteria for distinguishing the true from the false or the proved from the unproved. And this remains a danger no matter what our age or sophistication.[34] The combination of secrecy and power in all adults, but especially in the teacher/assessor, provides the conditions of later student cynicism and disrespect for both intellectual and moral standards. The *credibility* of the test and what the test is meant to represent—namely, adult standards—is at stake, and it is dependent upon the test maker's willingness to justify the test to takers and users and to accept questions and criticisms about the test's value from test takers.

In the absence of a mutually respectful relationship with the assessor, the student cannot appreciate the assessor's reasoning and internalize it. Paradoxically, according to Piaget, *even if I respect the assessors, the authority figures, I am unlikely to effectively uphold their standards*. Intellectual and moral standards, no matter how often "taught" or invoked by Authority, remain "external to the subject's conscience, [and] do not really transform his conduct" if the standards are adhered to by force of character or inequality of relationship only.[35] Is this not what we see and decry in students all the time? Is it not just what we mean when we talk about "thoughtless" student behavior? Listen also to our students' language as they discuss work turned back: "My English teacher doesn't like my writing," for example, conveys the view that the subjective and hazy desires of Authority are being taught, not clear standards. But if the correctness of answers and arguments is justified only because Authority says so, careless and thoughtless work is the inevitable result. So is passivity: students learn by such mysterious and one-way assessment that they cannot reframe questions, reject questions as inappropriate, challenge their premises, or propose a better way to prove their mastery. The moral and political harm is significant. Too many students learn to just "give them what they want" and to accept or acquiesce in bogus but "authoritative" judgments.

Excellence is not (and must never seem to be) about satisfying elders and their inexplicit and apparently contradictory tastes. Excellence must be seen as the meeting of known and justifiable demands—what (objectively) *works*.[36] We then come to understand a lesson that is morally as well as intellectually empowering: the judge too is subservient to intelligent principles.

Testing the Test for Excessive Secrecy

Various overarching criteria for the use of secrecy in testing surface out of this discussion. Students are entitled to the minimal secrecy necessary. To be autonomous (meaning, as the word's roots imply, "self-regulating"), students must confront assessment procedures that are *maximally transparent*—that is, procedures that are rehearsable to the greatest possible extent without compromising the test's validity or the purposes of school learning.[37] Students have a right to practice well the central tasks and standards by which they will be judged. In addition, they must know that what the assessor wants is of objective intellectual worth. Students must always be able to verify the aptness of the test, the score, and the assessor's use of secrecy. Otherwise, they resort to merely guessing or to calculating what the teacher "wants." In practice, this requirement of verification means that testing contexts should always make it possible for the student to self-assess and self-adjust before it is too late. It might even extend to building in a self-assessment of performance as part of the test. (A preponderance of inaccurate self-assessments might well cast doubt on the validity of the test.)

High-stakes and large-scale secret tests (such as the SATs), in which the only apparent judges are electronic, threaten intellectual integrity even further, because they heap harmful suspicion on human judgment. Intellectual freedom is thus threatened by secure

BEST COPY AVAILABLE

testing in the same way that it is threatened by political rule that emphasizes ritual and secrecy above the consent of the governed. Is it cavalier to suggest that the centrally mandated and secret testing of students in this country parallels the centrally planned and secret-dominated political systems now collapsing throughout Eastern Europe? I think not, for in Poland and Czechoslovakia, workers and citizens learned to do what many students and school districts regularly end up doing here: obey the letter of the law only, do no more and no less than the state directives require, see no incentives—indeed, see many disincentives—to set and honor more authentic standards. Nor do the makers and buyers of tests ever have to provide anything more than a public relations answer to critical questions raised by skeptics.

Ultimately, we must recognize that there is a fundamental cultural inequality at work in sustaining this harmful overreliance on test security. Pervasive test security is not necessary. One's place in the educational and professional hierarchy matters more than some theoretical concern with validity and feasibility in test construction. Our tolerance for incomparable results is suddenly much higher when the adults and powerful institutions are being assessed.[1] Much test security is required merely to enable the test to be uniform, simple, cheap, and widely used. But personal and institutional idiosyncrasies, complex quality-control procedures, and local autonomy—properly considered virtues of higher education, the professions, and the marketplace—are all compromised by generic secure tests. Continued reliance on secure proxy tests is thus possible only when schools and students have inadequate power to protest their overuse. Perhaps more to the point, such moral inequality is counterproductive (a theme to be picked up in Chapters Seven and Eight in an analysis of the differences between authentic standards and mere standardization).

A clear alternative approach is already being practiced in various U.S. school districts and in countries other than our own: Return control of assessment to teachers or their representatives; make the assessment more credible, contextual, and fair. Develop districtwide and statewide "task banks" of *exemplary* assessment tasks—tests worth teaching to and emulating in one's design. Develop clear and validated district performance standards through

benchmarking and a community-discussion process. Develop oversight and audit functions to ensure the soundness of the system. When a pool of questions and tasks that is large enough to prevent corruptive teaching has been developed, make those questions and tasks public, as if to say to teachers and students alike, "These represent the range of tasks that you must be good at to graduate. Now that you know this, the responsibility becomes yours [just as responsibility is assigned to the athlete, drafting student, artist, and debater]. We will coach you until you know how to self-assess and perform well enough that we become obsolete as 'teachers' and upholders of standards. Then you will have become, quite properly, our colleague." No mystery or secret there—which is just as it should be if our aim is to empower students and teachers instead of merely to check up on them and keep them in check.

Notes

1. This chapter is a substantially revised version of "Secure Tests, Insecure Test Takers," in J. Schwartz and K. A. Viator, eds., *The Prices of Secrecy: The Social, Intellectual and Psychological Costs of Testing in America*, a Report to the Ford Foundation (Cambridge, Mass.: Educational Technology Center, Harvard Graduate School of Education, 1990).
2. R. W. Highland, *A Guide for Use in Performance Testing in Air Force Technical Schools* (Lowry Air Force Base, Colo.: Armament Systems Personnel Research Laboratory, 1955).
3. R. Fitzpatrick and E. J. Morrison, "Performance and Product Evaluation," in F. L. Finch, ed., *Educational Performance Assessment* (Chicago: Riverside/Houghton Mifflin, [1971] 1991), p. 127.
4. S. Bok, *Secrets: On the Ethics of Concealment and Revelation* (New York: Random House, 1983), p. 25.
5. See the "Environmental Impact Statement" developed by the national watchdog group Fair Test for an example of how such a process might work.
6. As translated by N. Lewis, *Hans Andersen's Fairy Tales: A New Translation* (London: Puffin Books, 1981), p. 42.
7. Consider, by contrast, the recent manifesto behind Britain's new

BEST COPY AVAILABLE

national assessment design (which will rely heavily on classroom-based, teacher-overseen assessment): "A system which was not closely linked to normal classroom assessments and which did not demand the professional skills and commitment of teachers might be less expensive and simpler to implement, but would be indefensible in that it could set in opposition learning, teaching, and assessment." From Department of Education and Science, Assessment Performance Unit, *Mathematical Development*, Secondary Survey Report 1 (London: Her Majesty's Stationery Office, 1980), para. 48, p. 220.

8. The vaunted college admissions test, the SAT, is a lovely example of what happens when secrecy shrouds the testing process. It is *not* an achievement test, but it is used to whip districts and states into a frenzy about school achievement—despite explicit warnings in ETS material not to do so and the obvious fact that such tests depend heavily on socially constrained views of "intelligence" (and thus socioeconomic status). And who recalls that the SAT was developed as an *aptitude* test, for *equity* reasons—namely, to find students with potential in the hinterlands whose achievement might be limited by local schooling?
9. See R. Stiggins, "Assessment Literacy," *Phi Delta Kappan* 72 (1991): 534-539.
10. Educators in other countries derisively refer to our fetish for multiple-choice tests as "the American solution" to educational problems.
11. Bok, *Secrets*, p. 27.
12. Ibid., p. 110.
13. M. Mead, as quoted in Bok, *Secrets*, pp. 244-245.
14. M. Foucault, *Discipline and Punish*. (New York: Vintage Books, 1979), p. 35.
15. Ibid.
16. Ibid., p. 227.
17. S. Berryman, "Sending Clear Signals to Schools and Labor Markets," in J. Schwartz and K. A. Viator, eds., *The Prices of Secrecy: The Social, Intellectual, and Psychological Costs of Testing in America* (Cambridge, Mass.: Educational Technology Center, Harvard Graduate School of Education, 1990), p. 43.
18. See Bok, *Secrets*, chap. 4.
19. See a discussion of this legal history in R. A. Berk, ed., *Performance Assessment Methods and Applications* (Baltimore, Md.: Johns Hopkins University Press, 1986).
20. See Chairman of the New York State Senate Higher Education Committee, *Truth in Testing: A Study in Educational Reform*, revised report (Albany, N.Y.: New York State Senate, 1984); A. Strenio, *The Testing Trap* (New York: Rawson, Wade, 1981); and G. Blumenstyk, "Federal Court Ruling that Public Disclosure of Test Violates Copyright Law Seen as a Blow to 'Truth in Testing' Movement," *Chronicle of Higher Education*, Jan. 31, 1990.
21. American Psychological Association, "Standards for Educational and Psychological Testing" (Washington, D.C.: American Psychological Association, 1985), p. 2.
22. This argument against unobtrusive testing is ironic in light of the new British assessment of students; unobtrusiveness is now an explicit aim of formal British testing programs. See Chapter Five, on incentives, for related discussion. See also J. Millman and J. Greene, "The Specification and Development of Tests of Achievement and Ability," in R. Linn, ed., *Educational Measurement*, 3rd ed. (New York: American Council on Education/Macmillan, 1989), for another view on the questionable ethics of test unobtrusiveness.
23. W. Haney and G. Madaus, "The Evolution of Ethical and Technical Standards for Testing," in R. K. Hambleton and J. N. Zaal, eds., *Advances in Educational and Psychological Testing: Theory and Applications* (Norwell, Mass.: Kluwer, 1991).
24. Schwartz and Viator, eds. *The Prices of Secrecy*.
25. Ibid., p. 115.
26. Ibid., p. 116.
27. For an excellent discussion of authentic assessment and how to maximize the beneficial impact of tests on schooling—"systemic" validity—see J. R. Fredriksen and A. Collins, "A Systems

BEST COPY AVAILABLE

Approach to Educational Testing," *Educational Researcher 18*, 1989, 27-32.

28. Note that tests are not valid or invalid, in other words; inferences about results on tests are valid or invalid. Our failure to keep this point in mind is both common and unfortunate. See Chapter Seven for a further discussion of validity.
29. J. Millman and J. Greene, "The Specification and Development of Tests of Achievement and Ability," p. 347.
30. J. Piaget, *The Language and Thought of the Child*, trans. M. Gabain (New York: New American Library, [1932] 1965), p. 405.
31. Most educators do not seem to realize that establishing test validity is not merely a technical matter, based on technical rules and criteria. Validity can be established only by showing that the test results can be used to derive broader inferences about the value of the test scores vis-à-vis the wider world. Test makers need to *justify* their claims, ironically enough, by appealing to empirical evidence, not just statistics—something few school officials and policy makers require.
32. J. Piaget, *The Moral Judgment of the Child* (New York: Macmillan, [1932] 1965), p. 196.
33. Piaget, *The Moral Judgment of the Child*, pp. 137, 190. Piaget also suggests that schools reinforce the harmful effects of this constraint by overrelying on unverifiable didacticism, where "words are spoken with authority"; instead, adults should "discuss things on an equal footing and collaborate" with the child and lead the child to respect for both answers and their rational grounds (p. 194).
34. Piaget's sobering words were somehow lost on Kohlberg and others, who became convinced that intellectual development about moral matters was synonymous with irreversible moral progress. See C. Gilligan and G. Wiggins, "The Origins of Morality in Early Childhood Relationships," in J. Kagan and S. Lamb, *The Emergence of Morality in Young Children* (Chicago: University of Chicago Press, 1987).
35. Piaget, *The Moral Judgment of the Child*, p. 62.
36. The reader should not hear this in a utilitarian or mechanical sense. Art teachers often talk this way: Does the student's product work in its own way to accomplish the artist's end? And the AP art portfolio assesses for such things as a sense of focus, style, and personal direction.

37. See Fredriksen and Collins, "A Systems Approach to Educational Testing."
38. The exception, admissions tests for graduate schools, is a reminder that the function of most tests is to expedite the work of the gatekeepers at a "higher" point in the system. Contrary to popular belief, competitive college and graduate admissions offices run offices of *rejection*, not admission: a candidate's low test scores enable the readers of many folders to shorten their work considerably by reducing the "maybe" pile (barring counterevidence from transcripts or references.)

BEST COPY AVAILABLE

4 Testing and Tact

It was once a commonplace that tact is an essential capacity for teachers—*tact* in the sense described by William James as the "ingenuity to tell us what definite things to say and do when the pupil is before us . . . that tact for the concrete situation."[1] Socrates had this tact: each dialogue in which he participates is unique (as true dialogue must be); each move on his part derives from the unanticipatable response or objection from the other discussant.

It is just this tact that teaches students to respect the teacher in return. The teacher who is interested enough to ferret out what we are thinking, who persists in trying to understand our unexpected and opaque answers, and who sometimes understands our answers better than we do ultimately provides powerful incentive for us to develop the self-discipline required to refine our thinking and our answers. We develop a respect shown for intellectual standards based on the respect for our answers and the rules of evidence modeled by the standard-bearer. As Bruner once put it, "It requires a sensitive teacher to distinguish an intuitive mistake—an interestingly wrong leap—from a stupid or ignorant mistake, and it requires a teacher who can give approval and correction simultaneously."[2]

What, then, of the *assessor's* obligations? Is it not reasonable to ask that a tester be tactful? Well, of *course* testers ought to be tactful, you might say, but how is this even an issue in paper-and-pencil testing? As we have seen in previous chapters, a test is insensitive and unresponsive (by design) to answers deserving follow-up

or to questions that properly demand personalization by the answerer. Thus a tester, without intending to do so, treats the student as an object while tacitly demanding, "Answer *my* question, *my* way. You have one try in which to do it; and you get no opportunity to rethink your answer based on feedback, because you'll get no feedback." In a relationship, such a demand would be viewed (properly) as tactless, insensitive, overbearing. I would argue that it is equally inappropriate in a test situation. The student may lose respect for someone who behaves this way (whether in a relationship or in a classroom); the student may lose respect for and interest in what is assessed in such a manner. Perhaps the efficacy of such a method, and not just its ethics, is thrown into question.

This is not a hopelessly romantic idea. Binet and Piaget were adamant about the assessor's need to adapt the climate and questioning to the child's needs and to unanticipated results. Invalid inferences result when an assessment fails to honor Binet's caution that "the attitude to be taken in regard to the child is delicate: first of all, there must be good will; one must try to remain in touch with him; excite his attention, his self respect."[3] Second, there is an *unavoidable* ambiguity to mere answers, even if blackened in with finality by a #2 pencil: we do not know why the student selected a correct or incorrect answer. A correct answer might be due to irrelevant or lucky thinking; an incorrect answer might hide a thoughtful and apt response that, if pursued, would change our judgment as to the student's knowledge. Tact is a moral description of the obligation to validate our test results by exploring the reasons for answers. Being tactful in the assessment or testing relationship itself, especially when one has a singular goal in mind for one's work, is far harder than it appears. But it is essential that we become more tactful in assessment if our aim is to learn what the student really knows and to induce the student to care to know and reveal more.

Primum non nocere is the Latin phrase for the key element of the Hippocratic Oath: First, do no harm. But obeying the oath in all clinical settings is more difficult than it seems, since harm is usually caused by the unseen gap between our intentions and our actual effects. Teachers do not set out to be boring; they too often are. Test makers do not set out to influence learning negatively; they often do. (Stanford University professor of education Lee Shulman

BEST COPY AVAILABLE

once told me, in a discussion about this problem and his ongoing research into teaching, that the better teachers can often be distinguished by a seemingly mundane trait: they more accurately describe what goes on in their classroom than do less successful teachers; in other words, they *see* that gap between intention and effect.) The inclination to rationalize what occurs is ever-present when our intent is noble.

Consider a cautionary tale from the history of medicine as a reminder of how difficult this "tact for the concrete situation" can be. We go back to a time when doctors were struggling to treat a horrible new kind of problem, the wounds inflicted from a then-novel technology of war—the gun. Doctors proposed a cure for gunshot wounds based on the plausible view that the bullet poisoned the body. It was necessary, from this perspective, to treat gunshot wounds the way doctors treated any serious infection: the body needed to "undergo a cleansing with boiling oil." As one historian of medicine has put it, "Not only was the theory erroneous, but the treatment was ferociously traumatic. The resultant pain was as intolerable as was the destruction of tissue . . . yet the therapeutic iniquity persisted, enforced by the wrong-headed dogma."[4]

It took a young barber-surgeon, unschooled enough to trust his senses, and the lucky depletion of oil at a critical moment in battle to overcome the "theory." Ambroise Paré, who became the personal surgeon to a general of Francois I, was compelled to develop an unusual method of treatment during the siege of Turin in 1537, when the oil habitually used to cleanse wounds ran out with the unexpectedly high casualties. He had the "inspiration to design a battlefield clinical experiment. . . . [T]his novice surgeon hit upon the idea of making up a bland soothing lotion." As he noted, "I was constrained to apply a digestive of yolkes of eggs, oil of roses, and turpentine. . . . [Rising early in fear of whether my own method would work] beyond my expectation I found those to whom I applied my medicine, to feele little paine, and their wounds without inflammation: the others to whom was used the boiling oil, I found them feverish, with great pain and swelling. And then I resolved with myself never so cruelly to burne poore men."[5]

It may seem like the worst kind of disrespect and exaggeration to imply that test makers and test users are scarring their students in similar ways. But testers, like doctors, are *inevitably* prone to lose their tact when they rely primarily on their technical expertise and theory.[6] How rarely we see the harm of uniform testing and its presumed practices, after all. The theory of testing—like the theory of cleansing gunshot wounds—is strong enough and sanctioned enough to blind us to its effects.

We might see this potential harm more clearly if we thought of the test maker (and test user) and the test taker as in a relationship. The question to ask, then, is whether or not the tester is respectful and tactful in this relationship. If assessment is not something we do *to* students but something we do *with* them, then considering what is good or ill for the student in the assessment relationship is an essential obligation. Is it not, for example, intellectually disrespectful when a test maker feels no obligation to make test tasks thought-provoking enough to be worthy of the students' time and engagement? And what if the tester seems not to care whether students regularly get right answers for the wrong reasons (or vice versa)? What of the routine use of test security, distracters, and other questionable practices that seem both demeaning and liable to undermine students' faith in the assessor? What if we found out that certain often-administered tests and formats were perceived by students or teachers as boring, onerous, or not consistent with the experienced curriculum? What if an indirect and decontextualized form of testing is always less engaging than a direct method of assessment but is nonetheless more common?

And what of the student's relationship to knowledge itself? What disrespect is implied when we offer only generic, decontextualized, set exercises with right and wrong answers, as opposed to problems that demand judgment in the use and adaptation of knowledge? What bad lessons about the intellectual world are provided by tactless procedures that permit no student questions, no reframing of questions, and no opportunity to defend an odd or unorthodox response? If students (and the teacher, in the case of externally constructed tests) see no value in a test (and thus decide not to give the test their all), then what is the test actually measuring? Do test takers and test users have a right to be part of the validation process, just as all parties in a relationship are entitled to a say when the conduct of the relationship affects them (some-

BEST COPY AVAILABLE

thing we acknowledge with even small children)? At the very least, we need to consider how the design of any test adversely or positively affects each student's conception of learning and its value, as well as the student's "aspirations, motivation, and self-concept" (in the words of Benjamin Bloom, George Madaus, and Thomas Hastings).[7]

In this chapter, I propose to address questions of this sort—questions that should make us increasingly uncomfortable with the inherently unresponsive nature of most tests. I am fully prepared for some readers to find such a discussion naive and idealistic. But the reader would do well to reflect upon the consequences of a system of intellectual assessment that perpetually misleads the student about what knowledge is, how it is validated, the role that dialogue plays in real-world appraisals, and what kind of know-how and initiative in problem solving is really wanted in the adult world. At the very least, the reader should consider the disincentives provided by a testing system in which only sanctioned, uniform answers are desired and in which the assessor is never obliged to honor the student's wish to clarify questions and to critique or agree to the worth of the test.

Students' Relationship to Knowledge as Defined by Testing

If the "doing" of science or history is inherently more engaging than the learning of other people's conclusions, what consequences does this have for assessment? Or, to put it negatively, what should we make of what Howard Gardner has noted—namely, that "one of the most objectionable, though seldom remarked upon, features of formal testing is the intrinsic dullness of the materials. How often does anyone get excited about a test or a particular item?"[8] I note in Chapter Five that John Goodlad found most students to be more inclined to find value and pleasure in the arts, physical education, and vocational courses than in the traditional educational core. He discovered that the school activities that students found most pleasurable were those that "involved them actively or in which they worked with others. These included making films, building or drawing things, making collections, interviewing people, acting things out, and carrying out projects."[9] And as I noted in Chapter One, Bloom described synthesis as the "production of a *unique* communication" that "bears the stamp of the person." Synthesis is the place in the Taxonomy "which most clearly provides for creative behavior on the part of the learner."[10] Work that involves this creative behavior is clearly of greater value to the learner.

There is no better way to consider the impact of the test tasks themselves on student engagement than to compare different tests of the same subject. Consider the mathematics questions presented in Exhibit 4.1. Who cannot see the inherently engaging qualities of the m&m's problem? In addition, it may do a more thorough job of assessing for the outcome in question: understanding of volume.

I have used the m&m's problem shown in Exhibit 4.1 with many middle and high school classes after the appropriate (though fairly traditional) units had been completed by their teachers. What was striking to the teachers observing the students work through the problem each time was the intense interest demonstrated by even the poorest-performing students. In fact, in two cases, the "worst" students were among the last to leave the room, and they were in the top group of performers! And one pair of "poor" students pleaded with their teacher for another piece of posterboard so that they might work with it after school. (The teacher's jaw dropped visibly.)

Contextual detail and even ambiguity heighten engagement, because one's judgment, style, and wit are tapped, not just one's textbook knowledge. Consider, for example, the case method used to teach law, medicine, and business. Anyone who has witnessed the use of the case method in action, as I have in problem-based-learning medical courses (where students are confronted with cases requiring diagnosis and given only a list of symptoms and readings gained in emergency room admissions), is immediately struck by the heightened student engagement generated. The richness of the case, the need for deft judgment and application of knowledge—these are the very things that provide each student with incentive to enter into and grapple with the case.

Why does it make sense to call this engaging of students an issue of tact? Because a simplistic, generic test item out of context is both alienating and unresponsive to the realities of diverse human intelligence. (Is it too farfetched to say that we have had a

BEST COPY AVAILABLE

Exhibit 4.1. Two Approaches to Testing in Math.

Standardized Test Questions on Volume

1. What is the volume of a cone that has a base area of 78 square centimeters and a height of 12 centimeters?
   a. 30 cm^3
   b. 312 cm^3
   c. 936 cm^3
   d. 2808 cm^3

2. A round and a square cylinder share the same height. Which has the greater volume?

A Multiday Performance Assessment On Volume

Background: Manufacturers naturally want to spend as little as possible, not only on the product, but on packing and shipping it to stores. They want to *minimize* the cost of production of their packaging, and they want to *maximize* the amount of what is packaged inside (to keep handling and postage costs down: the more individual packages you ship, the more it costs).

Setting: Imagine that your group of two or three people is one of many in the packing department responsible for m&m's candies. The manager of the shipping department has found that the cheapest material for shipping comes as a flat piece of rectangular paperboard (the piece of posterboard you will be given). She is asking each work group in the packing department to help solve this problem: *What completely closed container, built out of the given piece of posterboard, will hold the* largest volume *of m&m's for safe shipping?*

1. Prove, in a *convincing* written report to company executives, that both the *shape* and the *dimensions* of your group's container maximize the volume. In making your case, supply all important data and formulas. Your group will also be asked to make a three-minute oral report at the next staff meeting. Both reports will be judged for *accuracy, thoroughness,* and *persuasiveness.*
2. Build a model (or multiple models) out of the posterboard of the container shape and size that you think solves the problem. The models are *not* proof; they will *illustrate* the claims you offer in your report.

gender gap in mathematics because the testing, not just the teaching, is so decontextualized and thus arelational?) The reason for using moral language to discuss this issue is that technical demands may not require designers to sweat over these kinds of "aesthetic" and contextual details. After all, they might well gain valid results from more morally questionable methods. My aim is to alert test users to the fact that test designers invariably work at a distance from students, literally and relationally; they are too infrequently asked to consider the moral and intellectual quality of their relationship to the student, as mediated by the test. The tester's intellectual interests are not those of the student, and yet *both* sets of interests should be considered. And not just in terms of the "social" implications of testing: while I applaud the work of Samuel Messick, senior researcher at Educational Testing Service, and others in bringing "consequential validity" as an issue before the measurement community, a "social" perspective may be as tactless and impersonal as a "merely technical" perspective. We need to better consider the differences between students and the unanticipatable but apt answers of students; this is what I mean by *tact.*

In the case of the test tasks themselves, we need to consider the student's relationship to the domain being assessed—to knowledge itself—as operationalized in the test. Dewey was one of the first and most vociferous of modern thinkers to argue that learning must always consider this relationship—the student's relationship to ideas and adult insights. Like any human relationship, the student's relationship to knowledge can be authentic or inauthentic, nurturing or distancing. His criticism of textbook-driven syllabi and tests rested in part on the harm to motivation that he saw resulting from inauthentically presenting knowledge as a fixed and settled compendium instead of as methods of inquiry and their (always tentative) results.[11]

Testing that operationally defines mastery in terms of the student's ability to recognize sanctioned answers makes clear that the student's contributions as a *thinker* are unimportant. Since "the child knows perfectly well that the [tester] and all his fellow pupils have exactly the same facts and ideas . . . he is not giving them anything at all new." It is not too extreme to suggest that the student is then treated as (and made to feel like) a mere object in the

BEST COPY AVAILABLE

test relationship: "As a consequence of the absence of the materials and occupations which generate real problems, the pupil's problems are not his; or rather, they are his only as a pupil, not as a human being. . . . Relationship to subject matter is no longer direct. . . .[T]he occasions and material of thought are not found in the arithmetic or history or geography itself, but in skillfully adapting that material to the [tester's] requirements."[12]

Anyone with common sense or with experience in school can bear witness to the student as object. We naturally want to contribute, to produce, to perform, to make a difference, to use our wits and style; we naturally are made frustrated, passive, or cynical by doing work that does not appear to be in our interests and is barely personalized. And in terms of validity concerns, higher-order thinking typically requires performance challenges that truly engage, as was noted above. (A concern for the student's relationship to knowledge must, therefore, be viewed in part as a demand for a more robust and authentic approach to construct validity, as I argue in Chapter Seven.)

Given that direct testing is more engaging and educative than indirect testing, we need to ask a fundamental question: Is the student *entitled* to the most "direct" forms of measurement possible, irrespective of the technical and logistical hurdles? After all, there are grounds for worry about the link between the nature and import of a test and student motivation. Paul Burke, the author of a recent paper that was part of a major congressional study of national testing, argues that there is a good deal of anecdotal and circumstantial evidence to show that NAEP results are of questionable worth, because students have no perceived interest in the test and its results.[13] Regrettably, Burke suggests that the only solution is to avoid tests that are "penalty-free"—as if only threats motivate eighth- and twelfth-graders to care about their work.

The dilemma is further illustrated in a curious passage in Diane Ravitch and Chester Finn's book describing the first NAEP U.S. history and literature test. In the pilot of the test, NAEP had included the option "I don't know" among the test responses. Although Ravitch and Finn advocated keeping that option in the multiple-choices items, "in accordance with its usual procedure, NAEP did not include 'I don't know' as a potential answer. It is the

test maker's judgment that this response is somehow inappropriate, that it confuses the analysis of the test results."[14] But that is not the full story. As one of the ETS researchers who worked on the project told me, the "confusion" is due to the fact that too many kids choose "I don't know" to mean "I don't care" (since the NAEP test has no bearing on their school grades or future aspirations).

We can cite a happier example to show the positive intrinsic incentives possible in formal testing challenges. New York has now for three years had a hands-on test of basic science skills as part of the fourth-grade program evaluation. The test is given to all students and is composed of five activities organized as separate "stations" to which the students rotate for a fixed period of time until all stations have been completed.[15] Douglas Reynolds, director of science for the state, tells a story echoed by many fourth-grade teachers I have worked with in New York: he visited on test day one of the first schools at which the test was given, and as the test ended, three students excitedly crowded around the teacher to ask, "Can we do this again tomorrow?"

Issues of Respect: The Unexamined Ethics of the Assessment Relationship

Whatever our view about the assessor's obligation to provide authentic forms of assessment and to give a voice to the test takers and test users, there remain other ethical concerns about testing that have been ignored in most overly technical discussions. Whether we consider the secrecy described in Chapter Three or the universal reliance on test formats that do not permit the test taker to ask questions, consult resources, or justify an answer, we ought to consider the link between technical practices of questionable merit and their possible negative impact on student performance, motivation, and trust.

Distracters and Other Forms of Deceit

Consider, for example, our unending reliance on deception in testing. Every test that provides a full range of possible answers from which the student selects involves chicanery—legal trickery—on the

BEST COPY AVAILABLE

part of the test maker. The student must choose between answers that are (and are meant to be) similar in appearance, though only one is correct (or "best"). We rarely consider the morality and the efficacy of such a practice, however, despite the fact that the very term used to describe those other possible answers—the "distracters"—alerts us to the test maker's deliberate deception. (More modern texts describe distracters as "foils." Better public relations, perhaps, but the intent is the same.)

Just so we know what we are talking about, here are two multiple-choice questions from actual tests that provide plausible but incorrect answer options:

Using the metric system, you would measure the length of your classroom in

a. grams
b. kilometers
c. meters
d. liters

A requirement for a democratic government is that

a. the people must have a voice in government
b. government leaders must do what they think is best for the people
c. judges must run for election rather than be appointed
d. everyone must pay the same amount of taxes

In the first example, from a science test for middle school students, we can easily imagine that a student could be "distracted" by the word *kilometers*, since it contains the key word *meter*. (Would placing option b after option c change the number of students selecting it?) Though c is clearly the desired answer, "kilometers" is in fact acceptable—and might even be "best" in a particular context whose overall scale was immense. In the second question, the first two choices have been made almost equally enticing, given such phrases as "the people," "voice," and "is best for the people." But which answer is the "right" one?

Deception and ambiguity in assessment are not necessarily wrong. There are clearly times when the uncertainty generated in students by plausible, diverse answer options is essential for ferreting out real understanding.[16] But we *routinely* deceive young students by providing misleading answer options at their and our moral and intellectual peril. What is wanted are some guidelines for knowing when such deception is warranted and when it threatens students' respect for the assessor and, by implication, what is being assessed. (This demand is easily frustrated, however, by the secure and proprietary nature of such tests. Nonetheless, we must demand better professional and public oversight procedures and standards.)

Alas, we search in vain for any mention of the ethics of such deception in the psychometric literature. There is no mention of the problem of using distracters in the APA Standards, for example. In the most recent comprehensive analysis of large-scale test-construction principles and issues, distracter *analysis* is discussed ("Among high-scoring examinees, the keyed answer should be popular. . . . All foils should be chosen by a sufficient percentage of low-scoring examinees to warrant their use. . . . The discrimination value of the keyed answer should be positive, for the foils, negative"), but the *ethics* of the practice are not.[17]

Similarly, in most measurement textbooks, we find only technical advice on how to fashion the answer options: "Foils are designed to attract examinees who lack understanding of the content or who are unable to use the material in the desired way. Therefore, foils must be plausible responses for those examinees whose mastery of content is incomplete."[18] Leaving aside the obliqueness of the phrase "who are unable to use the material in the desired way," how are we to find and validate such foils? "The best way to create foils is to anticipate (predict) the kinds of misunderstandings or errors in knowledge that examinees are likely to experience." The authors even propose a field test for this: "A very objective method of creating foils for multiple-choice questions is to first write foils as completion items and try them out on students. If a wide range of student ability is represented in the field-test sample, a large number of plausible wrong answers will be obtained. These wrong answers will usually make good foils for multiple-choice versions of the items."[19] The authors note that "foils should *not* trick knowledgeable examinees into incorrect answers." They do not offer advice on how to determine whether that is occurring, however.

Another rule for composing effective distracters, found in an

BEST COPY AVAILABLE

old text on test construction, is that the item writer "should maximize the ratio of plausibility to correctness in his distracters."[20] Similarly, another theorist argues that "a portion of the art of item writing is the ability to conceive of distracters which are incorrect by any standard, and yet attractive to those not knowing the answer."[21] Still a third argues that "if an item distracter is obviously a wrong answer to everyone . . . this distracter probably should be replaced by a more plausible answer."[22]

Should we not be sure, before condoning trickery, that people are being tricked for the right reasons? If our reasons are in question, should we continue to assume that the use of such trickery is an appropriate tool if it has some technical value? Is the use of distracters so different from the use of inappropriately leading questions or inappropriately gained evidence in law? It is tactless and perhaps morally objectionable to assess student understanding of an issue on the basis of a solitary tricky question; we could easily prevent this with more explicit design standards for the use of such test strategies.

To "maximize the ratio of plausibility to correctness" is clearly easier said than done, anyway—particularly when we see how easy it is to mislead students *simply by virtue of the fact that authorities are proposing the answers and the answers are plausible*. Although Binet, the father of modern testing, used distracters, he warned us of their perhaps excessive power to deceive. He devoted one of the set of intelligence tests to what he called "suggestibility"—the force of a student's "judgment of a subject and his powers of resistance" to foils that were deliberately involved in the question. Each "test" was built upon deliberately disingenuous, impossible-to-honor instructions by the examiner: the student is asked to indicate the button when handed a thread, thimble, and cup; the instructor asks the student to find a "nitchevo" (or other nonsense word) in a picture; having already looked at pairs of unequal lines and designated the longer and shorter, the student is shown a pair of equal lines and asked "And here?" As Binet notes of this last prompt, "Led on by the former replies [the student] has a tendency, a required force," to succumb to what Binet calls a "snare."[23]

One wonders why Binet included such tests, then, particularly since he explicitly warns the reader that suggestibility is "by

no means a test of intelligence, because very many people of superior intelligence are susceptible to suggestion through distraction, timidity, and fear of doing wrong."[24] In fact, in a later version of the test, Binet stressed that the child "falls into the trap because he allows himself to follow the lead of habit . . . [often due to] heedlessness, lack of attention."[25]

In another test, he rethought his original use of distracters. To test the student's ability to think logically, he made use of various logical absurdities, introduced into the test without warning. Examples included, "When a man plays marbles, why is he often decorated" and "When two men quarrel, why is there often near them a yellow dog?" Here too he noted that "very intelligent people sometimes fall into the trap. . . . [T]imidity, diffidence, confidence, and habit each plays its part." Binet thus notes the change in the test procedure: "Now instead of imposing an absurdity, we warn the child that it will come, and we ask him to discover it and refute it"; the examiner asks the student, "What is silly in this sentence?" Binet notes that "this experiment usually interests the children by its novelty."[26]

Piaget repeatedly stressed that the most difficult thing for all interviewers to avoid doing is eliciting what he called "suggested convictions." Especially important is the seductive power of the questioner's choice of words: "If one carelessly uses a word that is unexpected to the child, one risks stimulating . . . reactions which might then be mistaken for spontaneous." For example, "in asking 'what makes the sun move?' the suggestion of an outside agent occurs at once, thus provoking the creation of a myth."[27]

There is even a danger in the very method of testing, if one poses uniform questions under uniform conditions. The "essential failure of the test method" is that it "falsifies the natural mental inclination of the subject or risks doing so. . . . [I]t may well be that a child would never put the question to itself in such a form." The only way to avoid such difficulties is to "vary the questions, to make counter-suggestions, in short, to give up all idea of a fixed questionnaire." Since the "real problem is to know how he frames the question to himself or if he frames it at all," the skill of the assessor "consists not in making him answer questions . . . but encouraging the flow of spontaneous tendencies."[28] Piaget had some advice for

BEST COPY AVAILABLE

those intending to use his methods to ascertain what the child *really* thinks: "[The questioning of a student] requires extremely delicate handling. . . . Play your part in a simple spirit and let the child feel a certain superiority."[29]

Consider how the following examples from a national standardized multiple-choice achievement test in mathematics violate Piaget's caution and advice:

1. In which numeral is there a 4 in both the hundred thousands and hundreds places?
   a. 140,426
   b. 410,642
   c. 416,402
   d. 414,602

2. How many faces does a cube have?
   a. 4
   b. 6
   c. 8
   d. 12

3. Which of these would you most likely measure in ounces?
   a. paint in a can
   b. gasoline in a car
   c. water in a bathtub
   d. medicine in a bottle

Leaving aside the apparent irrelevancy of such questions (these are clearly not questions that "one would likely pose to oneself" or even that would be posed in an arithmetic course), they (and others like them) are inherently "distracting."[30] And since "errors" might therefore be made for a variety of reasons having little to do with understanding of the concepts in question, it is unclear whether such questions can be adequately validated in a formal sense. What is wanted, in fact, are better guidelines for when distracters ought to be used and, when they are deemed apt, better guidelines for their validation.

Alan Shoenfeld has catalogued many "suggested convictions," one example of which, the army bus problem, I noted in Chapter Two. He also cites the work of a European researcher who came up with the following results to show how suggestion by the

tester and the habitual demands of the classroom can be stronger than the student's judgment and knowledge:

> Reusser gave students a number of similar problems, for example: "There are 125 sheep and 5 dogs in a flock. How old is the shepherd?" The following quote from a fourth grade student working the problem out loud speaks for itself: "125 + 5 = 130 . . . this is too big, and 125 - 5 = 120 is still too big . . . while . . . 125/5 = 25. That works. . . . I think the shepherd is 25 years old." Also, he asked 101 fourth and fifth grade students to work the following problem: "Yesterday 33 boats sailed into the port and 54 boats left it. Yesterday at noon there were 40 boats left in the port. How many boats were still in the port yesterday evening?" He reports that all but one of the 101 students produced a numerical solution to the problem, and that only one student complained that the problem was unsolvable.[31]

We know from the work of Stanley Milgram, Ellen Langer, and others in psychology that adults are all-too suggestible; they are capable of induced thoughtlessness when commanded by "authorities" or situations to do odd or inappropriate things—none of which bears any indication of what they might be shown to know if the situation were less intimidating or less suggestive of the tester's "obvious" insights.

Recognition of human suggestibility and avoidance of test deceit are more than issues of tact. The answers that result from testing built upon trickery and suggestion may well be untrustworthy—invalid for making the kinds of inferences we want to make. If the child is seduced into certain answers, if the child is not "stimulated to any effort of adaptation" and produces answers at random or as a kind of "romancing" (Piaget's term for the invention of an answer that the child does not really believe to be true), then what have we learned? This has long been the position held by teachers of young students, who often assert that standardized tests are misleading and harmful to children and to their own work

BEST COPY AVAILABLE

BEST COPY AVAILABLE

as teachers. Their claim deserves to be better investigated and considered as a matter for policy.

Ambiguity and Equity

Note that the problem of unresponsive testing has little to do with whether the test is multiple-choice. In fact, many open-ended and performance tests use tasks and criteria of questionable merit and give students little opportunity to justify inherently ambiguous responses. For example, in a pilot of an open-ended mathematics test for third-graders in one southern state, the student is given two pictures, one of a cube and one of a pyramid, and asked, "How are these figures alike and how are they different?" The following scoring rubric is used:

> 0 Does not address the task, or no correct comparisons
> 1 Makes at least one comparison, but possibly only one category (alike or different) used
> 2 Makes at least 3 correct comparisons, with at least one in each category based on some geometrical property
> 3 Makes at least two correct comparisons in each category that are clearly stated and based on some geometrical property; no incorrect comparisons

Never mind the numerous questions about validity and score reliability that can be asked about such a question (I, for one, cannot fathom what this task is measuring or how the rubric can be validated as corresponding to genuine qualitative differences in performance); never mind the unfair secrecy in the student not having access to the judge's material. What about the *arbitrariness* of the rubric? From the wording of the question, why would the nine-year-old student imagine that the judges might reward for anything but the quality of her response? What validity can there be to a rubric that in fact discriminates answers on the basis of the number of a student's ideas only, not their quality? Or what about the fact that the scorers marked the following answer by a nine-year-old as

wrong: "They are different because one is a square and one is a triangle"? That answer is correct if we go by the face of each shape (as the illustration invited: each face was highlighted).

We can understand the assessor's obligation to be more tactful in situations such as these by thinking about the legal principle of "equity." (I am using the word *equity* in its original philosophical meaning, which was then incorporated into the British and American legal systems in the courts of chancery or equity.) The idea is a commonsense one, captured in a surprisingly complex legal history: blanket laws are inherently unable to encompass the inevitable idiosyncratic cases to which we ought always to make "exceptions to the rule." Aristotle put it best: "The equitable is a correction of the law where it is defective owing to its universality."[32] A standardized test, unresponsive (by design) to the inevitable eccentricities of individual student thought and the inevitable ambiguities of the questions themselves, is thus intrinsically "inequitable."

Put in the context of testing, the principle of equity requires that we ensure that human judgment and the "test" of responsive dialogue are not overrun by an efficient mechanical scoring system. Tests that must be administered in a uniform, unidirectional way (and in which clarification of student answers is forbidden) are dangerously immune to the possibility that a student might legitimately require the opportunity to defend an unexpected or "incorrect" answer. How many times, for example, have *you* had to alter a judgment after your child, student, friend, or spouse explained an odd or "unacceptable" action or answer? Sometimes all a student needs is a rephrasing of the question to recall and use what he or she "knows." We rely on human judges in law, as in athletics, because the spirit of the law cannot be encompassed by the letter of the law; judgments cannot be reduced to algorithms. And because both questions and answers contain possible ambiguities, to gauge understanding we must explore an answer: there must be some possibility of dialogue between the assessor and the assessee to ensure that the student is fully examined.

The question that we must repeatedly ask is whether test makers are obligated to do a better job of probing a student's reasoning, particularly when the test format induces haste and efficiency in answering and when short answers are inherently

BEST COPY AVAILABLE

ambiguous. Consider, as evidence of the problem, the following example from the NAEP "Learning by Doing" science test that was piloted a few years ago.[33] On one of the tasks, students were given three sets of statistics that supposedly derived from a mini-Olympics that some children put on. The introductory text noted that the children "decided to make each event of the same importance." No other information provided bears on the question. The test presented the students with the results of three "events" as follows:

Child's Name	Frisbee Toss	Weight Lift	50-Yard Dash
Joe	40 yards	205 lbs.	9.5 sec.
Jose	30 yards	170 lbs.	8.0 secs.
Kim	45 yards	130 lbs.	9.0 secs.
Sarah	28 yards	120 lbs.	7.6 secs.
Zabi	48 yards	140 lbs.	8.3 secs.

The first question asks the student, "Who would be the all-around winner?"
The scoring manual gives the following criteria for judging an answer:

> Score 4 points for accurate ranking of the children's performance on each event and citing Zabi as the overall winner. Score 3 points for using a ranking approach . . . but misinterpreting performance on the dash event . . . and therefore, citing the wrong winner. Score 2 points for a response which cites an overall winner or a tie with an explanation that demonstrates some recognition that a quantitative means of comparison is needed. Score 1 point if the student makes a selection of an overall winner with an irrelevant or non-quantitative account or without providing an explanation. Score 0 for no response.

Makes sense, right? But now ask yourself how you would score the following response given by a third-grader, using the given criteria:

BEST COPY AVAILABLE

a. *Who would be the all-around winner?*
No one.

b. *Explain how you decided who would be the all-around winner. Be sure to show your work.*
No one is the all-around winner.

The NAEP scorer gave the answer a 1. We see why, if we consider only the criteria: the student failed to give an explanation and any numerically related calculations to support the answer, and the answer appears incorrect since there was a winner. (The student, of course, did not have the rubric, which might have changed her answer.)

Suppose we assume, however, just as Piaget and Binet always warned us to assume, that an answer is usually apt in the mind of a student. Could it be that the nine-year-old deliberately and correctly answered "No one," thinking that "all-around" meant "winner of *all* events"? And, if looked at in this way, could it not be that the child was *more* thoughtful than most in deliberately not taking the "bait" of part b (which would have caused the child to pause in his or her answer, presumably). The full-sentence answer in part b—remember, this is a nine-year-old—is revealing to me: it is emphatic, as if to say, "No, your question suggests that I *should* have found one all-around winner, but I won't be fooled; I stick to my answer that no one was." (Note, by the way, that in the scorer's manual the word *all-around* has been changed to *overall*.) The student did not, of course, "explain" the answer, but it is conceivable that the instruction was confusing, given that there was no "work" needed to determine that "no one" was the all-around winner. One quick follow-up question would have settled the matter as to what the student's answer meant.

How ironic to learn that many of the tasks borrowed by ETS from the British Assessment Performance Unit (APU) were differently scored in the American version than in the original version. (APU scoring is discussed in a 1980 report on assessment prepared by the British Department of Science & Education.[34]) The APU considered in its scores the students' postexperimental voiced reflections on their work, seeing those reflections as being of great im-

portance in determining what students understand. For example, in the six hands-on science tasks used to assess science skill in eleven-year-olds, the same question was asked each time, after the experimental work was over:[35]

> If you could do this experiment again, using the same things that you have here, would you do it in the same way or change some things that you did, to make the experiment better?

Responses were rated on a three-point scale as follows:

	Rating
shows awareness of variables which were not controlled, procedures which turned out to be ineffective, the need to repeat measurement, or criticizes other factors which are central, not peripheral, to the investigation	2
shows awareness of alternative procedures but unaware of particular deficiencies of those used (does not have very good reasons for suggesting changes)	1
uncritical of procedures used, can suggest neither deficiencies nor alternative procedures	0

At issue is a moral question with technical ramifications: to what extent is the tester responsible for ensuring that student answers are sufficiently explored or understood? To what extent is the assessor obliged to both excite the student's attention (so as to evoke all of what the student thinks and knows on a subject) and probe those answers that do not, on first glance, seem apt (but that may well be apt once we grasp the rationale at work)? And what does the constant failure to use such techniques teach the student about intellectual challenges and standards?

It is striking that in many of the APU test protocols, such as the science protocol cited above, the assessor is meant to probe, prompt, and even teach, if necessary, to be sure of the student's actual ability. In many of the APU tests, *the first answer (or lack of one) is not deemed a sufficient insight into the student's knowledge.*[36] Consider, for example, the following sections from the assessor's manual in a mathematics test for fifteen-year-olds involving the ideas of perimeter, area, and circumference:[37]

1. *Ask:* "What is the perimeter of a rectangle?" [write student answer]
2. Present sheet with rectangle ABCD. *Ask:* "Could you show me the perimeter of this rectangle?"
3. *Ask:* "How would you measure the perimeter of the rectangle?" *If necessary, prompt for full procedure.*

As the APU report noted, these multiple opportunities to reveal knowledge were necessary, because many of the answers to the initial question were incorrect or confusing, while later answers revealed greater understanding: "About half the students [gave a satisfactory definition or formula for finding the perimeter.] Nearly all the testers commented on the inability of students to express themselves. . . . Many testers commented that there was some confusion between area and perimeter so that some pupils gave responses such as the 'area around the outside.' Looking at the results for the next two questions [, however,] over 80% could indicate the perimeter on the diagram."

A similar sequence of questions further illustrates the prompting allowed:[38]

10. "Estimate the length of the circumference of this circle." (answer: approx. 22 cm)
11. "What could you use to check your estimate?" *(string, rulers, etc. are on the table) If no response, prompt for string:* "Might you do something with the string?"
13. "Is there any other method?" *If student does not suggest using* $C = \pi\delta$, *prompt with* "Would it help to measure the diameter of the circle?"

If pupil does not suggest using formula, prompt

BEST COPY AVAILABLE

"Would it help to measure the diameter of the circle?"

It is of note that there was a 3 to 6 percent improvement in performance when students were prompted in these various offhand ways. As we shall see in Chapter Seven, this view of understanding—the view that we sometimes "forget" what we "know"—is more sound psychologically and philosophically than merely counting the first answer to a question.

This constant ability by the assessor to intervene, cue, or prompt in the APU assessments does not corrupt the test results, however, as revealed by the scoring system:[39]

1 Unaided success
2 Success following one prompt from the tester
3 Success following a series of prompts
4 Teaching by the tester; prompts unsuccessful
5 An unsuccessful response; tester did not prompt or teach
6 An unsuccessful response despite prompting and teaching
7 Question not given
8 Unaided success where student corrected an unsuccessful attempt without help

Successful responses were combined into 2 larger categories called "unaided success" and "unaided plus aided success" with percentages given for each.

The problem is not merely sloppy item writing nor poor rubrics; it is the inherent ambiguities in questions and answers that require some dialogue or follow-up. Consider the preliminary findings of Walt Haney and Laurie Scott in a study about ambiguity in test items. In an attempt to determine whether standardized test questions have inherent ambiguities that lead to misjudgments on the part of the assessor, Haney and Scott interviewed students on their responses. They often found, when students gave reasons for their answers, that "mistakes" and "correct answers" were not as

they appeared. Consider the following example from a reading comprehension item:

> Jane Addams cared about people all her life. When she lived in Chicago, she saw many children who did not have a good place to play. They played in the streets. . . . Jane Addams was given a large lot. The lot had many old, unused houses on it. She had the houses torn down so that the lot would be empty.
>
> Jane Addams worked hard to clean up the lot. Then she bought swings and seesaws. This became the first playground in Chicago.
>
> Five comprehension questions were based on this passage. One of them, and the answer alternatives posed for it, was, "What did Jane Addams do with the swings and seesaws?
>
> a. cleaned them
> b. played with them
> c. had them put on the lot
> d. used them in the streets
>
> Six out of ten children marked the third answer, which was the intended correct answer. Two marked "played with them." . . . One girl marked the first answer alternative, . . . "'Cause she doesn't want the kids to go on dirty things."

Haney and Scott note that the last child's inference is quite reasonable, given the comments about Addams's concern for cleanliness. The trouble is that this was not a test of "inference" but of "comprehension" from the test maker's point of view!

Haney and Scott catalogue dozens of similar examples and conclude that "ambiguity in children's interactions with test items is a phenomenon which occurs in a significant number of cases."[40] They recommend, as have other testing experts, that test makers routinely converse at length with students to probe these validity questions more carefully. They also warn that traditional notions

BEST COPY AVAILABLE

how to establish test validity are open to question as a result of their findings.

In sum, tact is not an option or a romantic demand. Any intellectual assessment is interpersonal—sensitive to different meanings and to context. One vital reason to safeguard the teacher's role as primary assessor is that the most accurate and equitable evaluation depends on relationships over time between examiner and student.[9] The teacher is the only one who knows what the student habitually can or cannot do, who has the ability to design completely contextualized and curricularly articulated assessments, and who has the ability and tact to follow up on confusing, glib, or ambiguous answers. In this country, we are so enamored of efficient testing that we overlook the feasible in-class alternatives to impersonal testing in use around the world. That is one reason why the German *abitur* (containing essay and oral questions) for graduation from the *Gymnasium* is still designed and scored by classroom teachers (who submit two possible tests to a state board for approval)—standards, but without standardization.

At the very least, we need to establish guidelines for the design of tests to ensure that the student's ability to justify or clarify a response is maximized. Tact is about responsiveness, after all. The relationship between test taker and tester must be made more morally balanced. In that relationship, as in all relationships, both responsiveness and equality are enhanced when each side has a say in the discussion. We can take that literally—"Let's have more oral examining"—or we can take it figuratively—"Let's ensure that the student has adequate opportunity to clarify both questions and answers." Either way, let us at *least* ensure that we "first, do no harm" to the student by testing; let us devise more rigorous procedures for determining and for counteracting the costs of tactless testing.

Notes

1. W. James, *Talks to Teachers* (New York: W. W. Norton, [1899] 1958), p. 24. For a thorough and illuminating discussion of tact and education, see M. Van Manen, *The Tact of Teaching: The Meaning of Pedagogical Thoughtfulness* (New York: State University of New York Press, 1991).
2. J. Bruner, *The Process of Education* (Cambridge, Mass.: Harvard University Press, 1960/1977), p. 68.
3. A. Binet and T. Simon, "New Investigation upon the Measure of the Intellectual Level Among School Children," in *The Development of Intelligence in Children* (Salem, N.H.: Ayer, [1911] 1985), p. 295.
4. S. B. Nuland, "The Gentle Surgeon," in *Doctors: The Biography of Medicine* (New York: Vintage Books, 1988), p. 97.
5. Quoted in Nuland, *Doctors*, p. 98.
6. The aptness of the comparison between medical and educational theory and practice goes beyond this delicate relationship between client, practitioner, and knowledge. Consider, for example, the current work of the Carnegie Board on Professional Teaching Standards, which is attempting to professionalize education in the same way that Flexner (also Carnegie-commissioned) led to standards in medical certification at the turn of the century.
7. B. S. Bloom, G. F. Madaus, and J. T. Hastings, *Evaluation to Improve Learning* (New York: Mc-Graw Hill, 1981), p. 51.
8. H. Gardner, *The Unschooled Mind: How Children Think and How Schools Should Teach* (New York: Basic Books, 1991), p. 98.
9. J. I. Goodlad, *A Place Called School: Prospects for the Future* (New York: McGraw-Hill, 1984), pp. 114–115.
10. B. S. Bloom, ed., *Taxonomy of Educational Objectives*, Vol 1: *Cognitive Domain* (White Plains, N.Y.: Longman, 1956), pp. 168, 175. Serious would-be test designers would do well to reread the *text* of the Taxonomy, not just the appendix/list, as well as the follow-up handbook developed by Bloom, Madaus, and Hastings years later (*Evaluation to Improve Learning*).
11. As an indication of the failure to grasp Dewey's point, see P. H. Hirst, "The Logical and Psychological Aspects of Teaching a Subject," and D. W. Hamlyn, "The Logical and Psychological Aspects of Learning," both in R. S. Peters, ed.,

BEST COPY AVAILABLE

The Concept of Education (London: Routledge & Kegan Paul).

12. "There is next to no opportunity for each child to work out something specifically his own, which may contribute to the common stock, while he in turn participates in the productions of others. . . . All are set to do the same work and turn out the same results. The social spirit is not cultivated, in fact it gradually atrophies for lack of use." J. Dewey, "Moral Principles in Education," in J. A. Boydston, ed., *The Middle Works of John Dewey, 1899–1924* (Carbondale: Southern Illinois University Press, [1909] 1977), p. 275; *Democracy in Education* (New York: Macmillan, 1916), p. 156.

13. P. Burke, *You Can Lead Adolescents to a Test But You Can't Make Them Try* (Washington, D.C.: U.S. Office of Technology Assessment, U.S. Department of Commerce/National Technical Information Service, 1991).

14. D. R. Ravitch and C. E. Finn, Jr., *What Do Our 17-Year-Olds Know?* (New York: HarperCollins, 1987), pp. 41–42. It is also worth noting that the authors were dismayed to learn that open-ended questions of the essay and short-answer type were not going to be used: "We would prefer an essay examination that determined the depth of student understanding. We hope that testing agencies will soon develop additional ways to assess knowledge and not rely so exclusively on multiple-choice questions" (p. 21). And, NAEP personnel have, in fact, developed more open-ended and performance tasks in content-area tests since the book was written.

15. See R. Mitchell, *Testing for Learning* (New York: Free Press/Macmillan, 1992), for a full account of this and other new state hands-on tests.

16. See P. Elbow, *Embracing Contraries: Explorations in Learning and Teaching* (New York: Oxford University Press, 1986), for a thoughtful discussion of the difference between the "teaching" role and the "assessing" role and the necessity of detached and critical probing in the latter.

17. J. Millman and J. Greene, "The Specification and Development of Tests of Achievement and Ability," in R. Linn, ed., *Educational Measurement*, 3rd ed. (New York: American Council on Education/Macmillan, 1989), pp. 361–362.

18. G. H. Roid and T. M. Haladyna, *A Technology for Test-Item Writing* (Orlando, Fla.: Harcourt Brace Jovanovich, 1982), pp. 180–181.

19. Ibid., p. 105.

20. R. L. Ebel, *Measuring Educational Achievement* (Englewood Cliffs, N.J.: Prentice-Hall, 1965), p. 168. By the way, there appears to be confusion over how the word should be spelled: both *distraction* and *distracter* can be found in many books, and both spellings appear in this Ebel book.

21. A. Wessman, cited in W. Haney and L. Scott, "Talking with Children About Tests: An Exploratory Study of Test Item Ambiguity," in R. O. Freedle and R. P. Duran, eds., *Cognitive and Linguistic Analyses of Test Performance* (Norwood, N.J.: Ablex, 1987), p. 865.

22. S. Henrysson, cited in Haney and Scott, "Talking with Children About Tests," p. 865.

23. A. Binet and T. Simon, "New Methods for the Diagnosis of the Intellectual Level of Subnormals," in *The Development of Intelligence in Children*. (Salem, N.H.: Ayer, [1905] 1988), p. 57.

24. Ibid., pp. 56–57.

25. Binet and Simon, "New Investigation upon the Measure of the Intellectual Level Among School Children," p. 285.

26. A. Binet and T. Simon, "The Development of Intelligence in the Child," in *The Development of Intelligence in Children* (Salem, N.H.: Ayer, [1908] 1988), pp. 227–228.

27. J. Piaget, *The Child's Conception of the World*, trans. J. Tomlinson and A. Tomlinson (Totowa, N.J.: Rowman & Allanheld, [1929] 1988), pp. 10ff., 15 and 5.

28. Ibid., pp. 5–4.

29. J. Piaget, *The Language and Thought of the Child*, trans. M. Gabain (New York: New American Library, [1932] 1974), pp. 25–26.

30. The test company has the nerve to suggest that the first question is a valid test of the construct "concept of numbers."

31. As reported in A. H. Shoenfeld, "Problem Solving in Con-

BEST COPY AVAILABLE

tent(s)," in R. Charles and E. Silver, eds., *The Teaching and Assessing of Mathematical Problem Solving* (Reston, Va.: National Council of Teachers of Mathematics; Erlbaum, 1988), pp. 83–84.

52. Aristotle, "Nichomachean Ethics," in J. Barnes, ed., *The Complete Works of Aristotle* (Princeton, N.J.: Princeton University Press, 1984), 1157b pp. 25–30. There is a lengthy and fascinating history to the concept of "equity."
53. National Assessment of Educational Progress, *Learning by Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics* (Princeton, N.J.: Educational Testing Service, 1987).
54. From Department of Education and Science, Assessment Performance Unit, *Mathematical Development*, Secondary Survey Report 1 (London: Her Majesty's Stationery Office, 1980), pp. 105–108.
55. From Department of Education and Science, Assessment Performance Unit, *Science in Schools: Age 11, Report No. 1* (London: Her Majesty's Stationery Office, 1981), p. 119.
56. Similar work on a research scale is being done in this country as part of what is called "diagnostic achievement assessment." See R. E. Snow, "Progress in Measurement, Cognitive Science, and Technology That Can Change the Relation Between Instruction and Assessment," in Educational Testing Service, ed., *Assessment in the Service of Learning*, Proceedings of the 1987 ETS Invitational Conference (Princeton, N.J.: Educational Testing Service, 1988), and J. S. Brown and R. R. Burton, "Diagnostic Models for Procedural Bugs in Basic Mathematical Skills," *Cognitive Science*, 1978, 2, 155–192.
57. Department of Education and Science, *Mathematical Development*, pp. 105ff.
58. *Ibid.*, pp. 106–107.
59. Department of Education and Science, *Science in Schools*, pp. 80–81.
60. Haney and Scott, "Talking with Children About Tests," p. 861.
61. See the research in language arts, for example, cited by Constance Weaver, in which the abilities that the standardized tests

purport to measure differ from the actual development of those abilities—and in which local assessments better support the construct. C. Weaver, *Understanding Whole Language* (Portsmouth, N.H.: Heinemann Educational Books, 1990).

BEST COPY AVAILABLE

BIBLIOGRAPHY

Selected Abstracts from the ERIC Educational Resources Database

ERIC Clearinghouse on
Reading, English, and Communication
Indiana University
Bloomington, Indiana

198

BEST COPY AVAILABLE

How to Read an ERIC Abstract and Find Related Articles on this Subject

The ERIC educational resource database includes more than 800,000 bibliographic records. Educational resources listed in the ERIC database are of two types: EJ, journal (magazine) articles, which are easily found in most Education libraries, or through interlibrary loan; and ED, documents such as Master's theses, which are available at any library that has an ERIC microfiche collection. ED documents can also be ordered directly from ERIC Document Reproduction Service by using the form at the end of this bibliography section.

You may also wish to perform your own ERIC database search, to retrieve the most current information on your topic. This is easily done at any Education library; it may also be available to you online through your university computing system.

In the following bibliography, we have selected some recent relevant articles that you may wish to read for your further knowledge, or to use in a Distance Education Application/Research Project. ERIC abstracts are easy to read, once you are used to the system, which is detailed below.

Sample ERIC Abstract

Note that this abstract has an EJ accession number, which means that the work abstracted is a journal article.

ERIC Accession Number—identification number sequentially assigned to articles as they are processed

Clearinghouse Accession Number

EJ446919 EC606207

Article Title → **Family-Centered Techniques: Integrating Enablement into the IFSP Process.** Andrews, Mary A.; Andrews, James R. *Journal of Childhood Communication Disorders*, v15 n1 p41–46 1993

Author(s) → Journal Title

Reprint Availability → (Reprint: UMI)

Descriptive Note → Note: Theme Issue: Service Delivery to Infants and Toddlers: Current Perspectives.

Volume No., Issue No., Pages, Publication Date

ISSN: 0735-3170 ← ISSN (International Standard Serial Number)

Major and Minor Descriptors—subject terms found in the *Thesaurus of ERIC Descriptors* that characterize substantive content. Only the major terms (preceded by an asterisk) are printed in the Subject Index of *Current Index to Journals in Education (CIJE)*

Descriptors: Child Rearing; *Communication Disorders; *Early Intervention; *Family Involvement; Individual Development; Objectives; Parenting Skills; Skill Development; *Teamwork; Young Children

Identifiers: *Enabler Model; Family Needs; *Individualized Family Service Plans ← Major and Minor Identifiers—terms found in the *Identifier Authority List* that characterize proper names or concepts not yet represented by descriptors. Only the major terms (preceded by an asterisk) are printed in the Subject Index of *Current Index to Journals in Education*

Annotation

This article describes techniques used in a family-centered early intervention project that both assist in accomplishing the goals of the Individualized Family Service Plan process and create opportunities for families to display their present competencies and acquire new ones to meet the needs of their children with communication disorders. (Author/JDD)

Annotator's Initials →

Note: The format of an ERIC Journal Article resume will vary according to the source from which the database is accessed. The above format is from the printed index, *Current Index to Journals in Education*.

1/95

BEST COPY AVAILABLE

GETTING COPIES OF THE ITEMS DESCRIBED IN THE ERIC DATABASE:

The items described in the ERIC database have either an "ED" or an "EJ" number in the first field. About 98% of the ED items can be found in the ERIC Microfiche Collection. The ERIC Document Reproduction Service (EDRS) in Alexandria, Virginia can produce either microfiche or paper copies of these documents for you. Check the accompanying list of ERIC Price Codes for their current prices.

Alternatively, you may prefer to consult the ERIC Microfiche Collection yourself before choosing documents to copy. Over 600 libraries in the United States subscribe to this collection. To find out which libraries near you have it, you are welcome to call the ERIC Clearinghouse on Reading and Communication Skills at (812) 855-5847. Most such libraries have equipment on site for inexpensive production of paper copies from the fiche.

For those few ED-numbered items not found in the Microfiche Collection, check the availability (AV) field of the citation to get information about the author, publisher, or other distributor.

Items with an EJ number in the first field of the citation are journal articles. Due to copyright restrictions, ERIC cannot provide copies of these articles. Most large college or university libraries subscribe to the journals in which these articles were published, and the general public can read or copy the articles from their collections. Should you want copies of articles which appeared in journals not owned by your nearest university library, arrangements usually can be made via interlibrary loan; there frequently is a nominal charge for this, which is set by the lending library. If you are a faculty member, staff member, or student at the university, just ask at your library's reference desk.

For all other categories of users, most universities cannot provide interlibrary services. However, public libraries---which are there to serve all area residents---typically are hooked into statewide lending networks designed to ensure that all state residents have access to materials of interest. Ask your local public librarian about interlibrary loan policies, charges, etc.

There are also two professional reprint services which have obtained permission from some journals to sell article copies. These are University Microfilms International (Article Clearinghouse, 300 North Zeeb Road, Ann Arbor, Michigan 48106---(800) 732-0616), and the Institute for Scientific Information (Original Article Tear Sheet Service, 3501 Market Street, Philadelphia, Pennsylvania 19104---(800) 523-1850). At the time of this publication, UMI charged $10.75 per article regardless of length, and ISI charged $9.50 for the first ten pages, plus $2.00 for each additional ten pages or fraction thereof. However, please check with them for current prices before ordering.

The following abstracts on Standardized and Alternative Assessment are from the ERIC educational resources database

AN: EJ430424
AU: Watkinson,-Anne
TI: **Primarily Assessing.**
PY: 1991
JN: Education-in-Science; n141 p8-9 Jan 1991
AV: UMI
DER: Curriculum-Evaluation; Elementary-Education; Foreign-Countries; Science-Education; Standardized-Tests; Standards-
DEM: *Academic-Achievement; *Elementary-School-Science; *Evaluation-Methods; *Observation-; *Student-Evaluation
AB: The idea of assessment driving the curriculum and the importance of assessing students constantly instead of using only standardized tests are discussed. Informal observations used by teachers as a way of assessing student progress is emphasized. The United Kingdom's National Curriculum assessment program are described. (KR)

AN: EJ430344
AU: Perrone,-Vito
TI: **On Standardized Testing.**
PY: 1991
JN: Childhood-Education; v67 n3 p131-42 Spr 1991
AV: UMI
DER: Achievement-Tests; Elementary-Education; Performance-Tests; Position-Papers; Student-Development; Teacher-Made-Tests; Teacher-Student-Relationship; Teaching-Methods; Writing-Evaluation
DEM: *Educational-Testing; *Elementary-School-Students; *Standardized-Tests; *Student-Evaluation; *Test-Use
AB: Reviews the history and uses of standardized testing, and suggests alternative methods of assessment. States the belief that testing exerts undue pressure on children, provides no useful information on individual children, and limits educational possibilities. Discusses the Association for Childhood Education International position paper which states that all testing of young children in preschool and grades K-2 should cease. (BC)

AN: EJ408022
AU: Norris,-Stephen-P.
TI: **Can We Test Validly for Critical Thinking?**
PY: 1989
JN: Educational-Researcher; v18 n9 p21-26 Dec 1989
AV: UMI
NT: Theme issue with title, "Educational Assessment."
DER: Elementary-Secondary-Education; Evaluation-; Problem-Solving
DEM: *Credibility-; *Critical-Thinking; *Multiple-Choice-Tests; *Test-Validity; *Verbal-Tests
AB: Discusses the generalizability of critical thinking and the disposition to think critically. Argues that verbal reports of examinees' thinking on multiple-choice tests can explain the reasoning behind their answers and, thus, can be used to assess the inability to make credibility judgments. (FMW)

AN: EJ519194
AU: Shepard,-Lorrie-A.; Bliem,-Carribeth-L.
TI: **Parents' Thinking about Standardized Tests and Performance Assessments.**
PY: 1995
JN: Educational-Researcher; v24 n8 p25-32 Nov 1995
AV: UMI
DER: Elementary-Secondary-Education; Parent-Attitudes; Public-Opinion; Test-Construction; Test-Results
DEM: *Achievement-Tests; *Educational-Assessment; *Parents-; *Standardized-Tests; *Test-Use
AB: Parental attitudes about standardized tests and performance assessments were studied through interviews with 60 parents or parent dyads in schools involved in a project to develop new assessments and through questionnaires completed by parents from control schools. Parent approval of standardized tests did not imply disapproval of performance assessment. (SLD)

AN: EJ517092
TI: **Ready-Reference Folder.**
PY: 1995
JN: Learning; v24 n1 p39-42 Aug 1995
AV: UMI
DER: Elementary-Secondary-Education; Portfolios-Background-Materials
DEM: *Evaluation-Methods; *Portfolio-Assessment; *Standardized-Tests; *Student-Evaluation
AB: A tear-out folder provides information on alternative student assessment versus standardized testing, describing what alternative assessments do and do not consist of and involve. Information on student assessment statewide and nationwide is presented along with a discussion of the costs of alternative assessment. All information is presented in color-coded charts. (SM)

AN: ED389734
AU: Shepard,-Lorrie-A.; Bliem,-Caribeth-L.
TI: **An Analysis of Parent Opinions and Changes in Opinions Regarding Standardized Tests, Teacher's Information, and Performance Assessments.**
CS: National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
PY: 1995
NT: 63 p.; Portions of the report presented at the Annual Meeting of the American Educational Research Association (1993).
PR: EDRS Price - MF01/PC03 Plus Postage.
DER: Beliefs-; Elementary-Secondary-Education; Parent-Teacher-Conferences; Parent-Teacher-Cooperation; Questionnaires-; Report-Cards
DEM: *Educational-Assessment; *Parent-Attitudes; *Parents-; *Standardized-Tests; *Test-Use
AB: Parent opinions about standardized tests and performance assessments were examined systematically. Mutually exclusive but randomly equivalent stratified samples from schools participating in a study of performance assessment and control schools were used to measure change in parent opinion over time. Approximately one-third of parents (n=105) completed questionnaires at the beginning of the school year, one-third completed them at the end of the year (similar sample), and the remaining third supplied interview samples (n=33 and n=27, respectively). Results demonstrated that parents' favorable ratings of standardized national tests did not imply a preference for this type of educational assessment over other types of assessment for measuring student or school progress. Parents considered report cards, hearing from the teacher, and seeing graded samples of student work as more informative than standardized tests, and they wanted comparative information to measure their own child's progress. When parents had a chance to look at

performance assessments through the year, they endorsed their use for district purposes and preferred them for classroom use. Survey data like the Gallup Poll showing widespread approval of standardized tests should not be taken to mean that parents are opposed to other forms of assessment. Appendixes contain the parent questionnaire and the interview protocol. (Contains 3 figures, 17 tables, and 9 references.) (SLD)

AN: ED388484
AU: Estrin,-Elise-Trumbull; Nelson-Barber,-Sharon
TI: **Issues in Cross-Cultural Assessment: American Indian and Alaska Native Students. Knowledge Brief, Number Twelve.**
CS: Far West Lab. for Educational Research and Development, San Francisco, Calif.; Native Education Initiative of the Regional Educational Labs.
PY: 1995
NT: 9 p.
PR: EDRS Price - MF01/PC01 Plus Postage.
DER: American-Indians; Classroom-Communication; Cultural-Context; Culture-Conflict; Educational-Environment; Educational-Strategies; Elementary-Secondary-Education; Epistemology-; Evaluation-Problems
DEM: *Alaska-Natives; *American-Indian-Education; *Cultural-Differences; *Culturally-Relevant-Education; *Instructional-Effectiveness; *Student-Evaluation
AB: This brief focuses on assessment issues for American Indian and Alaska Native (collectively referred to as "Native") students, as well as other pedagogical issues related to improved teaching and educational outcomes. Although traditional Native educational strategies emphasize cooperation, experiential learning, and reflection, Native students continue to be at a disadvantage in the classroom. The reasons lie in several intersecting realities: troubled historical relations between tribes and the federal government affecting Native schooling, ongoing educational practices that ignore or devalue cultural ways of knowing, and the dearth of American Indian and Alaska Native teachers. Understanding the school performance of Native students requires a sociocultural perspective that takes into account differences between community and school in social and cultural context, the unconscious nature of these contexts, effects on student learning and organization of knowledge, and implications for effective instructional styles and student evaluation. Despite supportive federal legislation, a repertoire of culturally specific instruments to assess Native student performance does not exist. Standardized norm-referenced tests present such difficulties as inappropriate content, time pressures, reliance on verbal information, basic premises of multiple-choice testing, and alien nature of formal on-demand testing. Indeed, achievement tests can be seen as merely indices of the student's acculturation to Western cultural knowledge and conventions for displaying knowledge. More culturally responsive assessment incorporates content reflecting local contexts and experiences, uses procedures that reflect local ways of thinking and learning, and provides students with options. Other concerns related to the question of whose standards are appropriate, proper interpretation and use of test data, and the value of alternative assessments. (SV)

AN: ED386707
AU: Holland,-Kathleen, Ed.; And-Others
TI: **Alternative Perspectives in Assessing Children's Language and Literacy.**
PY: 1994
AV: Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648 (cloth: ISBN-0-89391-864-4, $49.50; paperback: ISBN-0-89391-914-4, $24.50).
NT: 235 p.
PR: Document Not Available from EDRS.
DER: Elementary-Secondary-Education; Ethnography-; Evaluation-Methods; Language-Arts; Language-Skills; Sociolinguistics-
DEM: *Child-Language; *Literacy-; *Reader-Response; *Student-Evaluation; *Writing-Evaluation
AB: Suggesting many ways (not just one way) to understand what children are doing with language and literacy, this book presents essays that address the need for alternative perspectives as well as anthropological, socio-psycholinguistic, and reader response perspectives in assessing children's language and literacy. After "Introduction: What Is an Alternative?" by editors, David Bloome, Kathleen Holland, and Judith Solsken, part 1, which discusses the need for alternative perspectives, includes essays: (1) "Language, Culture and the Implications of Assessment" (Karla Holloway); and (2) "Towards an Alternative View of Writing Assessment" (Loren S. Barritt). Essays in the second part, which discusses anthropological perspectives, are: (3) "The Language of Testing: An Ethnographic-Sociolinguistic Perspective in Standardized Tests" (Catherine Emihovich); (4) "You Can't Get There from Here" (David Bloome); and (5) "Assessing Students as Members of a Literate Community" (Beth Gildin Watrous and Jerri Willett). A discussion follows: "Assessment in My World" (Maryann Jennings). Essays in the third part, which discusses socio-linguistic perspectives, are: (6) "Alternative Language Assessment: Communicating Naturally with Students in Assessment Contexts" (Helen B. Slaughter); (7) Assessing the Written Language Abilities of Beginning Writers" (Jo-Anne R. Wilson Keenan); and (8) "Looking at Their Own Words: Students' Assessment of Their Own Writing" (Susan Benedict). A discussion follows: "Making Assessment a Process" (Nina Tepper and Rocio Costa). Essays in the fourth part on reader response perspectives are: (9) "Children's Response to Literature: Isn't It about Time We Said Good-Bye to Book Reports and Literal Oral Book Discussions?" (Kathleen Holland and Susan Lehr); (10) "Assessing Literary Understanding through Oral Language" (Joanne M. Golden); and (11) "Children's Group Discussions of Literature: Fertile Ground for Informal Oral Language Assessment" (Lenore Carlisle). A discussion follows: "Unleashing the Potential of Children's Responses to Literature" (Leslie Shaw, Deborah G. Jacque, and Cheryl L. Taylor). (RS)

AN: EJ505556
AU: Schommer,-Marlene
TI: **Voices in Education on Authentic Assessment.**
PY: 1995
JN: Mid-Western-Educational-Researcher; v8 n2 p13-14 Spr 1995
DER: Costs-; Educational-Practices; Educational-Testing; Elementary-Secondary-Education; National-Competency-Tests
DEM: *Educational-Change; *Evaluation-Methods; *Program-Implementation; *Student-Evaluation; *Teacher-Attitudes
AB: Six prominent educators discuss barriers to achieving quality authentic assessment and the feasibility of implementing authentic assessment on a national basis. Points out the importance of using multiple assessments, the lack of a definition of authentic assessment, problems in developing quality assessments, and the immense cost of implementing authentic assessment nationally. (LP)

BEST COPY AVAILABLE

AN: ED384610
AU: Galloway,-Dan; Schwartz,-Wendell
TI: **Designing More Effective Grouping Practices at the High School Level.**
PY: 1994
NT: 30 p.; Paper presented at the Annual Meeting of the Association for Supervision and Curriculum Development (49th, Chicago, IL, March 19-22, 1994).
PR: EDRS Price - MF01/PC02 Plus Postage.
DER: At-Risk-Persons; Cooperative-Learning; Educational-Diagnosis; Educational-Improvement; Heterogeneous-Grouping; High-Schools; Homogeneous-Grouping; School-Restructuring; Self-Concept; Student-Motivation; Teacher-Expectations-of-Students; Tutorial-Programs
DEM: *Educational-Change; *Grouping-Instructional-Purposes; *Labeling-of-Persons; *Student-Evaluation; *Student-Placement
AB: Efforts at one high school to reconsider its practices of ability grouping and explore alternative assessment and grouping practices are described. Assessment of the schools' practices found that students in lower ability groups had a less stimulating curriculum, fewer positive role models, lower motivation, lower expectations for themselves, and worked with teachers who also held lower expectations for them. When mobility did take place between ability levels, it was more often downward than upward. The use of national standardized placement tests was replaced by teacher-designed, criterion-referenced assessment tools, resulting in significantly different balances of placements. A pilot program was launched to replace a remedial composition course with participation in regular level classes supplemented by ongoing lunch hour tutoring in composition, resulting in improved grades for participants. The success of this program led the school to eliminate lower ability levels in other content areas, and to modify curriculum in remaining lower-level courses. A variety of modifications were implemented to support heterogenous grouping, including expanded use of cooperative learning and classroom workshops. Staff development is seen as essential to the future of these modifications. (PB)

AN: ED382670
AU: Burstein,-Leigh
TI: **Performance-Based Assessment for Accountability Purposes: Taking the Plunge and Assessing the Consequences.**
CS: National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
PY: 1994
NT: 15 p.; A version of a paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 1991).
PR: EDRS Price - MF01/PC01 Plus Postage.
DER: Costs-; Educational-Change; Educational-Planning; Psychometrics-; Test-Construction; Test-Items
DEM: *Accountability-; *Educational-Assessment; *Multiple-Choice-Tests; *Standardized-Tests; *Testing-Problems; *Test-Use
AB: Issues in alternative assessment for accountability purposes are discussed. Most new forms of performance assessment are linked in the literature, but all alternative forms of assessment do not have the same attributes in terms of technical and feasibility criteria. Tradeoffs in the validity of inferences that can be drawn from alternative assessments and typical multiple-choice assessments must be assessed in light of the costs of performance assessment. There are problems to solve before implementing performance assessment including a host of technical and practical problems in terms of its use in large-scale assessment. Curriculum forces and the educational policy community are committed to shifting from assessment solely through multiple-choice to assessments that represent more important knowledge and skills closer to the attributes of interest. The question is not whether we will have performance assessment, but rather when. The way to move to performance assessment for accountability purposes is to tackle change in stages, planning for implementation and including planning for reporting results. Performance assessment is going to be a part of the response to the new American pluralism and to the increasingly global society. (Contains 12 references.)

AN: ED380786
AU: Stallman,-Anne-C.; And-Others
TI: **Alternative Approaches to Vocabulary Assessment. Technical Report No. 607.**
CS: Center for the Study of Reading, Urbana, IL.
PY: 1995
NT: 20 p.
PR: EDRS Price - MF01/PC01 Plus Postage.
DER: Academic-Achievement; Concurrent-Validity; Intermediate-Grades; Reading-Research; Reliability-; Standardized-Tests; Test-Validity
DEM: *Vocabulary-Development
AB: Interviews with children about their knowledge of a set of words was used to examine the concurrent validity of three paper-and-pencil measures of knowledge of these words--a standardized vocabulary test and two experimenter-designed tests. One experimenter-designed test, the Levels test, had three multiple-choice items per word that targeted three different levels of word knowledge. The other was a forced-choice contexts test with five items per word, each requiring a decision about whether the word was used appropriately in the context. Subjects were 50 students from two heterogeneously grouped fifth-grade classrooms in a midwestern school district. All three paper-and-pencil measures showed acceptable levels of reliability. When subjects were used as the unit of analysis, the interview was more highly correlated with the standardized test and the Levels test than with the Contexts test. When the word was used as the unit of analysis, the interview correlated more highly with the Contexts and the Levels test than with the standardized test. These results are interpreted as indicating that standardized measures are more effective at discriminating among students upon the basis of their overall ability, but less accurate as measures of how much the students know about particular words. The Contexts test has the advantages of the highest reliability of the three measures, as well as the greatest instructional validity. (Contains 25 references and 7 tables of data.) (Author/RS)

AN: ED380508
AU: Davis,-Wesley
TI: **Alternative Assessment: Facts and Opinions.**
CS: Florida Educational Research Council, Inc., Sanibel.
PY: 1994
JN: Florida-Educational-Research-Council-Research-Bulletin; v25 n4 p1-32 Sum 1994
AV: Florida Educational Research Council, Inc., P.O. Box 506, Sanibel, FL 33957 ($4; annual subscription, $15; 10% discount on 5 or more copies).
NT: 34 p.
PR: EDRS Price - MF01/PC02 Plus Postage.
DER: Educational-Change; Educational-Improvement; Elementary-Secondary-Education; Literature-Reviews;

Norm-Referenced-Tests; Standardized-Tests; Teacher-Role; Test-Use
DEM: *Cost-Effectiveness; *Educational-Assessment; *Evaluation-Methods; *Opinions-; *Student-Evaluation; *Test-Construction
AB: An attempt is made to separate facts from opinions based on review of a representative sample of contemporary writings on alternative assessment. A summary listing of 15 statements perceived to be factual is offered, followed by opinions of the author. These items cover: (1) the historical background and origins of alternative assessments; (2) their current intent, focus, and emphasis; (3) their technical problems and limitations; (4) the potential impact for change these procedures may have on instruction and student-teacher relationships; (5) other possible consequences of changes; (6) the expanded role of teachers in implementation; (7) the most significant contribution alternative assessment might make for students; and (8) projected cost factors. It is suggested that changing the instructional process for the better may well be the major contribution of alternative assessment. Large-scale norm-referenced standardized tests are here to stay, and cost factors may mean that alternative assessments are most useful in the individual classroom. One table summarizes facts about alternative assessment. (Contains 44 references.) (SLD)

AN: ED380483
AU: Jones,-Russell-W.
TI: **Performance and Alternative Assessment Techniques: Meeting the Challenge of Alternative Evaluation Strategies.**
PY: 1994
NT: 29 p.; Paper presented at the International Conference on Educational Evaluation and Assessment (2nd, Pretoria, Republic of South Africa, July 1994).
PR: EDRS Price - MF01/PC02 Plus Postage.
DER: Educational-Trends; Evaluation-Methods; Knowledge-Level; Portfolios-Background-Materials; Test-Format
DEM: *Educational-Assessment; *Multiple-Choice-Tests; *Portfolio-Assessment; *Test-Construction; *Testing-
AB: One of the most influential contemporary trends in educational evaluation in the United States is the move away from traditional testing methods toward "authentic assessments," which are designed to measure student performance of skills, abilities, and knowledge directly. While there is no consensus as to precisely what constitutes authentic assessment, there is general agreement that it incorporates: (1) emphasis on examinee performance, assessing not only what the examinee knows, but what the examinee can do; (2) use of direct methods of assessment; (3) inclusion of a high degree of realism; and (4) activities for which there may be no one correct answer, in a simulation of realism. The primary distinction between traditional testing methods and authentic assessment is the choice of question format. Certain segments of the educational community have called for a move from the multiple-choice format to question formats considered to reflect higher-order cognitive processes more accurately. This has resulted in the development or adoption of a broad range of formats, including standardized patient, audio-visual context setting, computer-based problem solving, multiple choice with justification, latent image, performance, and portfolio. Two figures illustrate the discussion. (Contains 30 references.) (Author/SLD)

AN: ED378187
AU: Foucar-Szocki,-Diane
TI: **Becoming Assessors: Authentic Assessment for Authentic Instruction. A Report of the Blue Ridge Assessment Project, a Collaborative Effort of the Albemarle, Fluvanna, Greene, Harrisonburg, Orange, and Rockingham Schools.**
CS: Albemarle County Schools, Charlottesville, Va.
PY: 1994
AV: Frank Morgan, Curriculum Development and Research, Albemarle County Schools, 401 McIntire Rd., Charlottesville, VA 22901.
NT: 435 p.; Portions of section 4 (resources) contain illegible print.
PR: EDRS Price - MF01/PC18 Plus Postage.
DER: Check-Lists; Elementary-Education; Inservice-Teacher-Education; Performance-Tests; Simulation-; Skill-Analysis; Student-Projects
DEM: *Evaluation-Methods; *Portfolio-Assessment; *Student-Evaluation; *Test-Construction
AB: This report describes educators' collaborative self-directed experiences in learning about and producing alternative forms of assessment for use in elementary-level classrooms. The report analyzes the Virginia Standards of Learning and local curricula in English/Language Arts, Mathematics, Social Studies, and Science/Health in grades 2-4, to provide common ground for the assessments. The assessments themselves are then presented, including checklists, portfolios, performance tasks, product assessments, projects, and simulations. Assessments in language arts cover oral communication, research and reporting skills, reading, writing, letter writing, and creative writing. Assessments in mathematics focus on problem solving, probability and statistics, data analysis, conservation, patterns, economics, money, measurement, geometry, and graphing. Social studies assessments address economics, earth care, Powhatan Indians, Tidewater region, and cooperative learning. Science/health assessments examine investigative skills and observation skills. A section titled "Voices" seasons the report with teachers' statements about learning and individual change in their professional contexts and their lives. Another section offers recommendations to enhance teacher learning and change. A resources section presents a glossary, and items for use as overhead transparencies or handouts. Appendices include a project plan, assessment item checklist, a model of the purposes of assessment, and a participant list. Contains 55 references. (JDD)

AN: EJ492908
AU: Hoerr,-Thomas-R.
TI: **How the New City School Applies the Multiple Intelligences.**
PY: 1994
JN: Educational-Leadership; v52 n3 p29-33 Nov 1994
AV: UMI
DER: Elementary-Education
DEM: *Curriculum-Development; *Intelligence-; *Portfolios-Background-Materials; *Professional-Development; *Theory-Practice-Relationship
AB: Describes a Saint Louis elementary school's successful application of Howard Gardner's multiple intelligences theory. What began as a discussion of the nature of intelligence has resulted in a revised curriculum, varied instructional techniques, alternative assessment (using a combination of portfolios, progress reports, profiles, demonstrations of understanding, and standardized tests), improved professional development for teachers, and new ways to communicate with parents. (MLH)

AN: EJ488411
AU: Charlesworth,-Rosalind; And-Others
TI: **Research on the Effects of Group Standardized Testing on Instruction, Pupils, and Teachers: New Directions for Policy.**
PY: 1994
JN: Early-Education-and-Development; v5 n3 p195-212 Jul 1994
DER: Evaluation-Methods; Student-Evaluation; Test-Coaching; Test-Wiseness
DEM: *Curriculum-Problems; *Group-Testing; *Standardized-Tests; *Test-Results; *Test-Use
AB: Since the late 1960s, curriculum and instruction have increasingly become "test driven." Research documents the negative effects of standardized testing on instruction, students, and teachers. At their best, tests reflect students' ability to take tests. Research supports the need for change in testing policies. Massive testing must be eliminated, and more authentic and performance-based evaluation should be instituted. (TM)

AN: ED374433
AU: Calfee,-Robert-C.
TI: **Ahead to the Past: Assessing Student Achievement in Writing. Occasional Paper No. 39.**
CS: National Center for the Study of Writing and Literacy, Berkeley, CA.
PY: 1994
NT: 14 p.
PR: EDRS Price - MF01/PC01 Plus Postage.
DER: Elementary-Secondary-Education; Higher-Education; Parent-School-Relationship
DEM: *Portfolios-Background-Materials; *Student-Evaluation; *Writing-Achievement; *Writing-Evaluation
AB: In the mid-1980s, partly through new developments in curriculum and instruction, a new assessment option has arisen under the banner of alternative assessment. The basic idea appears in several guises: "authentic assessment" (implying that standardized tests are not authentic), "performance tests," and "portfolios." The goal of writing portfolios is to provide an opportunity for a richer, more authentic assessment of their achievements, to show their potential given adequate time and resources. A survey of a broad array of portfolio practices around the country finds that: (1) the portfolio approach is energizing the professional standing of classroom teachers; (2) respondents showed a distaste for evaluation; and (3) teachers had little concern about technical matters like validity and reliability. Individual teachers interpret the portfolio concept quite differently in different settings. If portfolios are taken seriously, most students react seriously. One way to connect parents and schools is to place the students in a central role through the portfolio. Barriers to alternative assessments are substantial: time, money, motivation, and institutional support. The greatest hope for realizing the promise of portfolios may spring from the local school and the classroom teacher. Two caveats need to be observed: assessment practice and policies should be consistent for all teachers in a given school; and the audience and purpose for the assessment need to be established. Educators have made great strides during the past 50 years--the portfolio concept is but one example. (RS)

AN: ED373082
AU: Estrin,-Elise-Trumbull
TI: **Alternative Assessment: Issues in Language, Culture, and Equity. Knowledge Brief Number 11.**
CS: Far West Lab. for Educational Research and Development, San Francisco, Calif.
PY: 1993
NT: 9 p.
PR: EDRS Price - MF01/PC01 Plus Postage.
DER: Context-Effect; Elementary-Secondary-Education; Minority-Groups; Norm-Referenced-Tests; Performance-; Portfolios-Background-Materials; Social-Influences; Teaching-Methods; Test-Construction; Test-Validity; Track-System-Education
DEM: *Cultural-Differences; *Educational-Assessment; *Equal-Education; *Language-Proficiency; *Test-Use
AB: Alternative assessments, also called authentic or performance assessments, have first the common notion of a meaningful performance or product. The use of such alternatives with students belonging to nondominant language and cultural groups is of interest because of the hope that this type of assessment can show what these students can really do and because of the fear that the inequities that are associated with traditional norm-referenced tests will recur. Negative consequences of traditional forms of assessment for students from nondominant cultural and language groups are well documented. Equity issues in the public debate on any assessment are consequential validity, gatekeeping, tracking, and the opportunity to learn. The social context of assessment must be acknowledged in the design of alternative assessments. The alternative-assessment approach is compatible with a constructivist view of learners. Some suggestions are offered for a culturally responsive pedagogy and assessment design. A particular focus is on portfolio use in the pluralistic classroom. Three tables and one figure illustrate the discussion. (Contains 22 references.) (SLD)

ORDER FORM

ERIC DOCUMENT REPRODUCTION SERVICE - EDRS

Operated by: Cincinnati Bell Information Systems (CBIS) Federal Inc.
7420 Fullerton Road · Suite 110
Springfield, Virginia 22153-2852

Please allow 3 to 5 business days for EDRS to process your order and at least one week for shipping.

TELEPHONE ORDERS: Call 1-800-443-ERIC or (703) 440-1400
FAX ORDERS: (703) 440-1408
(24 hours per day, 7 days per week)

Address Information

Date: _____

Name: _____

Organization: _____

Ship To: _____

_____

City: _____ State: _____

County: _____ Zip _____

Tax Exempt: Yes No (Please Circle)

Tax Exemption Number: _____

Telephone: _____

Method of Payment

☐ Check or Money Order (U.S. Dollars) $_____

☐ Purchase Order (Over $50.00-Domestic Only) #_____

☐ Charge My Deposit Account #_____

Charge My ☐ Visa ☐ MasterCard

Account No. _____ Expir. Date _____

Signature: _____

(Required to Validate All Deposit Account and Credit Card Orders)

ORDERING INSTRUCTIONS

- Enter 6-Digit ED Number
- Enter Number of Pages in Document
- Enter Document Price from Price Schedule Below
- Enter Quantity, Specifying Microfiche (MF) or Paper Copy (PC)
- Enter Total Price per Document
- Enter Shipping Charges - See Instructions on Reverse

ED Number	No. of Pages	Document Price	Quantity MF	Quantity PC	Total Price
		Subtotal			
		Shipping			
		Tax			
		TOTAL COST			

ERIC DOCUMENT PRICE SCHEDULE**
Based on Number of Pages in Original Document

Paper Copy (PC) Price Code	No. of Pages	Price	Microfiche (MF) Price Code	No. of Pages	Price
PC01	1 - 25	$ 3.30	MF01	1 - 480	$1.18
PC02	26 - 50	$ 6.60	MF02	481 - 576	$1.43
PC03	51 - 75	$ 9.90	MF03	577 - 672	$1.68
PC04	76 - 100	$13.20	MF04	673 - 768	$1.93
PC05	101 - 125	$16.50	MF05	769 - 864	$2.18
PC06	126 - 150	$19.80	MF06	865 - 960	$2.43
PC07	151 - 175	$23.10	MF07	961 - 1056	$2.68
Each Additional 25 Pages or Fraction = $3.30			Each Additional 96 Pages (1 Microfiche) or Fraction = $.25		
PRICES ARE PER DOCUMENT					

****EDRS prices are good through December 31, 1993.**

Prices will change January 1 of each year.

All residents of AL, AZ, CA, CO, DC, FL, GA, IL, KY, LA, MA, MD, MO, NJ, NY, OH, PA, TX, UT, VA, and WA must add state and local taxes (subject to change).

GST will be collected on orders shipped to Canada.

EXPEDITED DOCUMENT DELIVERY

You may request expedited shipment by:

- USPS Express Mail
- UPS Next Day Air
- Federal Express
- FAX Transmission of Your Document

Shipping or FAX transmission charges will be added to the cost of the document(s) by EDRS.

Effective January 1, 1993

SHIPPING INFORMATION

Please consult appropriate rate chart.
UPS will not deliver to a P.O. Box address.

DOMESTIC: ALL ORDERS ARE SHIPPED AS FOLLOWS, UNLESS OTHERWISE SPECIFIED:

- All Paper Copy (PC) orders are shipped via UPS
- All Microfiche (MF) orders over 81 microfiche are shipped via UPS
- All Microfiche (MF) orders under 81 microfiche are shipped via USPS 1st Class

UPS rates as shown are based on the Zone furthest from Springfield, VA. Your shipping charges should not exceed these rates.

PLEASE NOTE: SHIPPING COSTS CAN CHANGE WITHOUT NOTICE

UPS RATE CHART
Shipping Charges should not exceed the following:

1 lb.	2 lbs.	3 lbs.	4 lbs.	5 lbs.
81-160 MF or 1-75 PC (Pages) $3.52	161-330 MF or 76-150 PC (Pages) $4.13	331-500 MF or 151-225 PC (Pages) $4.50	501-670 MF or 226-300 PC (Pages) $4.78	671-840 MF or 301-375 PC (Pages) $4.99
6 lbs.	**7 lbs.**	**8 lbs.**	**9 lbs.**	**10 lbs.**
841-1010 MF or 376-450 PC (Pages) $5.13	1011-1180 MF or 451-525 PC (Pages) $5.35	1181-1350 MF or 526-600 PC (Pages) $5.71	1351-1520 MF or 601-675 PC (Pages) $6.12	1521-1690 MF or 676-750 PC (Pages) $6.53

USPS FIRST CLASS RATE CHART

1-7 Microfiche	8-19 Microfiche	20-30 Microfiche	31-42 Microfiche	43-54 Microfiche	55-67 Microfiche	68-80 Microfiche
$.52	$.75	$.98	$1.21	$1.44	$1.67	$1.90

FOREIGN:

- Based on International Postage Rates in effect
- Allow 160 Microfiche or 75 Paper Copy pages per pound
- Specify exact mail classification desired

DEPOSIT ACCOUNTS

Customers who have a continuing need for ERIC Documents may open a Deposit Account by depositing a minimum of $300.00. Once an account is opened, ERIC Documents will be sent upon request, and the account charged for the actual cost and postage. A statement of the account will be furnished with each order.

STANDING ORDER SUBSCRIPTION ACCOUNTS

Subscription Orders for documents in the monthly issues of Resources in Education (RIE) are available on microfiche from EDRS. The microfiche are furnished on diazo film and without protective envelopes at $0.110 per microfiche. If you prefer silver halide film, the cost is $0.235 per microfiche, and each microfiche is inserted into an acid-free protective envelope. Prices are good through December 31, 1993, and do not include shipping charges. A Standing Order Account may be opened by depositing $2,300.00 or submitting an executed purchase order. All orders placed from outside the domestic U.S. must be prepaid. The cost of each issue and shipping will be charged against the account. A monthly statement of the account will be furnished.

BACK COLLECTIONS

Back collections of documents in all issues of RIE since 1966 are available on microfiche at a unit price of $0.141 per microfiche. The collections are furnished on diazo film without envelopes. Prices are good through December 31, 1993, and do not include shipping charges and applicable taxes. For pricing information, write or call toll-free 1-800-443-ERIC.

GENERAL INFORMATION

1. PAPER COPY (PC)

   A Paper Copy is a xerographic reproduction, on paper, from microfiche of the original document. Each paper copy has a Vellum Bristol cover to identify and protect the document.

2. PAYMENT

   The prices set forth herein do not include any sales, use, excise, or similar taxes that may apply to the sale of microfiche or paper copy to the customer. The cost of such taxes, if any, shall be borne by the customer.

   For all orders that are not prepaid and require an invoice, payment shall be made net thirty (30) days from the date of the invoice. Please make checks or money orders payable to CBIS (must be in U.S. funds and payable on a U.S. bank).

3. REPRODUCTION

   Permission to further reproduce a copyrighted document provided hereunder must be obtained from the copyright holder, usually noted on the front or back of the title page of the copyrighted document.

4. QUALITY

   CBIS Federal will only replace products returned because of reproduction defects or incompleteness caused by EDRS.